On the Foundations of Adversarial Single-Class Classification

Ran El-Yaniv RANI@CS.TECHNION.AC.IL
Department of Computer Science
Technion – Israel Institute of Technology
Technion, Israel 32000

Mordechai Nisenson MOTINIS@IL.IBM.COM
IBM Research – Haifa
Haifa University Campus, Mount Carmel, 31905 Haifa, Israel

Abstract

Motivated by authentication, intrusion and spam detection applications we consider single-class classification (SCC) as a two-person game between the learner and an adversary. In this game the learner has a sample from a target distribution and the goal is to construct a classifier capable of distinguishing observations from the target distribution from observations emitted from an unknown other distribution. The ideal SCC classifier must guarantee a given tolerance for the false-positive error (false alarm rate) while minimizing the false negative error (intruder pass rate). Viewing SCC as a two-person zero-sum game we identify both deterministic and randomized optimal classification strategies for different game variants. We demonstrate that randomized classification can provide a significant advantage. In the deterministic setting we show how to reduce SCC to two-class classification where in the two-class problem the other class is a synthetically generated distribution. We provide an efficient and practical algorithm for constructing and solving the two class problem. The algorithm distinguishes low density regions of the target distribution and is shown to be consistent.

1. Introduction

In *Single-Class Classification (SCC)* the learner observes a training set of sampled instances from one *target distribution*. The goal is to create a classifier that can distinguish instances emitted from distributions other than the target distribution and unknown to the learner during training. This SCC problem can model many applications such as intrusion, fault and novelty detection. For example, in an instance of an intrusion detection problem (see e.g., Nisenson, Yariv, El-Yaniv, & Meir, 2003), the goal is to create a classifier that can distinguish 'legal' users from intruders based on behaviometric or biometric patterns. This classifier can then be used to guard against illegal attempts to gain access into protected systems or regions.

Single-class classification (also termed one-class classification) has been receiving considerable research attention in the machine learning and pattern recognition communities. For example, only the survey papers (Markou & Singh, 2003a, 2003b; Hodge & Austin, 2004) cite, altogether, over 100 SCC papers. Most SCC works implicitly assume that a good solution can be achieved by identifying low density regions of the target distribution and then, the objective is to reject sub-domains of low density. Thus, the main consideration in previous SCC studies has been *statistical*: how can a prescribed false positive rate be guaranteed given a finite sample from the target distribution.

The proposed approaches are typically *generative* or *discriminative*. Generative solutions range from full density estimation (Bishop, 1994), to partial density estimation such as quantile estimation (G. Lanckriet, Ghaoui, & Jordan, 2002), level set estimation (Ben-David & Lindenbaum, 1995; Steinwart, Hush, & Scovel, 2005) or local density estimation (Breunig, Kriegel, Ng, & Sander, 2000). In discriminative methods one attempts to generate a decision boundary appropriately enclosing the high density regions of the training set (Yu, 2005). In addition to such constructions, there are many empirical studies of the proposed solutions. Nevertheless, it appears that the area suffers from a lack of theoretical contributions and principled (empirical) comparative studies of the proposed solutions.

Motivated mainly by intrusion detection applications, in this paper we examine the SCC problem from an adversarial viewpoint where an adversary selects the attacking distribution. We begin by abstracting away the statistical estimation component of the problem by considering a setting where the learner has a very large sample from the target distribution. This setting is modeled by assuming that the learning algorithm has precise knowledge of the target distribution. While this assumption would render almost the entire body of SCC literature superfluous, it turns out that a significant and non-trivial *decision-theoretic* component of the adversarial SCC problem remains – one that has so far been overlooked. For a discrete version of the SCC problem we provide an in depth analysis of adversarial SCC and identify optimal strategies for variants of the problem depending on whether or not the learner can play a randomized strategy and on various constraints on the adversary. As a consequence of this analysis, it can be demonstrated that a randomized learner strategy can be superior on average to standard deterministic classification. For an infinitely continuous version of this game we provide a simple and consistent SCC algorithm that implements the standard low-density rejection by reducing the SCC problem to two-class soft classification.

The body of this paper contains the principal results that are simpler to present. The appendices contain some of the more technical proofs. to the presented results. An earlier version of this work containing a subset of the results was presented at NIPS (El-Yaniv & Nisenson, 2006). Extensions to this work can be found in the thesis of (Nisenson, 2010).

2. Problem Formulation

We define the adversarial *single-class classification (SCC)* problem as a two-person zero-sum game between the *learner* and an *adversary*. The learner receives a training sample of examples from a *target distribution* P defined over some space Ω. On the basis of this training sample, the learner should select a rejection function $r : \Omega \to [0, 1]$, where for each $\omega \in \Omega$, $r(\omega)$ is the probability with which the learner will reject ω. On the basis of any knowledge of P and/or $r(\cdot)$, the adversary selects an *attacking distribution* Q, defined over Ω. Then, a new example is drawn from $\gamma P + (1 - \gamma)Q$, where $0 < \gamma < 1$, is a *switching probability* unknown to the learner.

The *rejection rate* of the learner, using a rejection function r, with respect to any distribution D (over Ω), is $\rho(r, D) \triangleq \mathbf{E}_D\{r(\omega)\}$. The two main quantities of interest here are the *false positive rate* (type I error) $\rho(r, P)$, and the *false negative rate* (type II error) $1-\rho(r, Q)$. Before the start of the game, the learner receives a tolerance parameter $0 < \delta < 1$, giving the maximally allowed false positive rate. A rejection function $r(\cdot)$ is *valid* if its false positive rate satisfies the constraint $\rho(r, P) \leq \delta$. A valid rejection function (strategy) is *optimal* if it guarantees the smallest false negative rate amongst all valid strategies.

This setting conveniently models various SCC applications and in particular, intrusion detection problems. For example, considering biometric authentication, the false alarm rate $\rho(r, P)$ is the rejection (failed authentication) rate of the legal users and $\rho(r, Q)$ is the rejection rate of intruders, which should be maximized.

Remark 1. Clearly, a dual SCC problem can be formulated where a sufficiently high intruder rejection rate must be guaranteed and the false alarm rate should be minimized. We briefly discuss this dual problem and its relation to the "primal" in Section 8. Other types of SCC problems can be considered where the loss is a function of the type I and type II errors. For example, one may be interested in minimizing a convex combination of these errors. Any such loss function can be handled using our definition and searching for the δ for which the SCC solution optimizes the desired loss function.

Our analysis begins by focusing on the Bayes decision theoretic version of the SCC problem in which the learner knows the target distribution P precisely. The problem is thus viewed as a two-person zero sum game where the payoff to the learner is $\rho(r, Q)$. The set $\mathcal{R}_\delta(P) \triangleq \{r : \rho(r, P) \leq \delta\}$ of valid rejection functions is the learner's strategy space. We denote by $\mathcal{Q}$ be the strategy space of the adversary, consisting of all allowable distributions Q that can be selected by the adversary.[1]

1. The game can be expressed in 'extensive form' (i.e., a game tree) where in the first move the learner selects a rejection function, followed by a chance move to determine the source (either P or Q) of the test example (with probability γ). In the case where Q is selected, the adversary chooses (randomly using Q) the test example. In this game the choice of Q depends on knowledge of P and $r(\cdot)$.

We are concerned with optimal learner strategies for game variants distinguished by the adversary's knowledge of the learner's strategy, P and/or of δ and by other limitations on $\mathcal{Q}$. We also distinguish a special type of this game, which we call the *hard setting* in which the learner is constrained to employ only deterministic reject functions; that is, $r : \Omega \to \{0, 1\}$, and such rejection functions are termed "hard." The more general game defined above (with "soft" functions) is called the *soft setting*. As far as we know, only the hard setting has been considered in the SCC literature thus far. The reason for considering soft rejection functions is that they can achieve significant advantage in terms of type II error reduction. Later on in Section 6.2.1 we numerically demonstrate such error reductions.

For any rejection function, the learner can reduce the type II error by rejecting more (i.e., by increasing $r(\cdot)$). Therefore, in the soft setting for an optimal $r(\cdot)$ we must have $\rho(r, P) = \delta$ (rather than $\rho(r, P) \leq \delta$). It follows that the switching parameter γ is immaterial to the selection of an optimal strategy.

Given an adversary strategy space, $\mathcal{Q}$, we define the set $\mathcal{R}^*_\delta(P)$ of optimal valid rejection functions as $\mathcal{R}^*_\delta \triangleq \{r \in \mathcal{R}_\delta(P) \ : \min_{Q\in\mathcal{Q}} \rho(r, Q) = \max_{r'\in\mathcal{R}_\delta(P)} \min_{Q'\in\mathcal{Q}} \rho(r', Q')\}$.[2] We note that $\mathcal{R}^*_\delta$ is never empty in the cases we consider.

3. Related Work

One-Class Classification is often given different names, depending on the desired use. For example, other common names include outlier detection, fault detection and novelty detection. Historically, one of the earliest works is due to Grubbs (1969) who considered in-sample outlier detection. Grubbs calculates a cut-off statistic for determining outliers in the 1-dimensional Gaussian case at the 5%, 2.5% and 1% significance levels within samples of various sizes. Minter (1975) appears to be the first to use the term "single-class classification". Minter starts from a fairly standard two-class approach, assuming that there is a class of interest (class 1) and a class of "others" (class $\emptyset$). Given the switching parameter γ (which is the a priori probability of class 1), Minter gives the rule to accept a point x iff $\gamma \Pr\{x|1\} \geq (1-\gamma) \Pr\{x|\emptyset\}$, which is equivalent to $\gamma \Pr\{x|1\} \geq \frac{1}{2} \Pr\{x\}$. It is assumed that both γ and $\Pr\{x\}$ are known or can be estimated from historical data, leaving the problem of estimating $\Pr\{x|1\}$ from the given sample. While, technically, only a sample from the class of interest is given, the additional assumptions make this a modified form of a two-class problem.[3] These are the earliest explicit works we have found. Note that statisticians have

2. For certain strategy spaces, $\mathcal{Q}$, it may be necessary to consider the infimum rather than the minimum. In such cases it may be necessary to replace '$Q \in \mathcal{Q}$' (in definitions, theorems, etc.) with '$Q \in cl(\mathcal{Q})$', where $cl(\mathcal{Q})$ is the closure of $\mathcal{Q}$.
3. This differs from more recent works, where γ and $\Pr\{x\}$ are assumed to be unknown (whereby the learner's knowledge is much more restricted), and the type I error is required to not exceed a bound, δ, which is the setting we use in this work.

long been considering the *two-sample problem*, which is similar but perhaps simpler. One can view the SCC problem as an extremely unbalanced instance of the two-sample problem that prevents using the standard statistical hypothesis testing techniques.

Since virtually all prior works on SCC that we have encountered deal with how to approximate a low-density rejection strategy given a set, $\{x_1, \ldots, x_n\}$, of training points, sampled from the class of interest, we will focus our review here on such methods.

We begin with discussing support, quantile and level-set estimation. Support estimation aims to estimate the support of a density p. In terms of outlier detection, the goal is clear: a point falling outside the estimated support is taken to be an outlier. One of the simpler methods, analyzed by Devroye and Wise (1980), is to estimate the support as $\hat{S}_n = \bigcup_{i=1}^{n} B(x_i, \epsilon_n)$, where $B(x, a)$ is a closed ball centered at x with radius a (i.e. $||x' - x|| \leq a$, for some norm $||\cdot||$), and ϵ_n is a (vanishing) sequence of smoothing parameters. In quantile estimation, the goal is to find a set $U(\beta)$ such that $\lambda(U(\beta)) = \inf_S\{\lambda(S) : P(S) > \beta\}$, where λ is a real valued function. For our purposes, we take λ as the Lebesgue measure, in which case the problem is also called *minimum volume estimation*. When $\beta = 0$ this becomes support estimation, and when $\beta = 1 - \delta$ this problem is the same as low-density rejection. In level-set estimation, the goal is to approximate the set $\mathcal{L}(t) = \{x : p(x) > t\}$ (or alternatively as $\{x : p(x) \geq t\}$). Of course, level-set estimation can be used for support estimation by taking $t = 0$ or by taking $t = t_n$ as a sequence which approaches zero (see Cuevas & Fraiman, 1997). Clearly, level-set estimation approximates the low-density rejection strategy when $P(\mathcal{L}(t)) = 1 - \delta$. A significant amount of prior SCC works have focused on minimum volume and level-set estimation. We distinguish between explicit and implicit methods, where explicit methods try to directly solve one of the problems, and implicit methods which use a heuristic which may or may not give the desired result. We note that whether the method is explicit or implicit is not necessarily an indicator of whether the underlying model is generative or discriminative, although there is a clear tendency for explicit methods to be generative. Transformations from the one-class setting to the two-class setting tend to be implicit and discriminative. We will consider minimum volume estimation approaches first and then look at various level-set estimation results. Finally we will examine other results, including transformations to the two-class setting.

Minimum volume estimation has been a favored approach at solving the SCC problem in the literature. This perhaps is due to two works which reused the popular Support Vector Machine (SVM, see Vapnik, 1998) from two-class classification problems. The earlier work (D. Tax & Duin, 1999) sought to fit the sample data inside a sphere of minimal radius, a solution they called the Support-Vector Data Description (SVDD). Specifically, given a sphere with center a and radius R, the error function to be minimized is $R^2 + C\sum_i \xi_i$, under the constraints $(x_i - a)^T(x_i - a) \leq R^2 + \xi_i$, where C is a regularization term which relates to

the type I error. Outliers in the sample data would lie on, or outside the sphere (and have $\xi_i > 0$). The kernel trick was then employed to allow for solving the problem in a higher dimensional feature space. They note that polynomial kernels do not result in small volumes in the input space, as points distant from the origin tend to have high error values. They found that Gaussian kernels worked well. The type I error can be estimated from the number of support vectors divided by the sample size, n, where the support vectors are the points lying on the sphere (i.e. they define the sphere's boundary). Changing the regularization parameter C, or the bandwidth parameter of the Gaussian kernel, can be used to control the trade-off between the volume of the sphere and the number of support vectors. In a follow up work, D. M. J. Tax and Duin (2001) show how samples from a uniform distribution can be used to optimize for both parameters simultaneously. The second work (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001) introduced what is commonly called the One-Class Support Vector Machine (OC-SVM). The technique used is that of a standard two-class SVM where the second class is the origin (in *feature space*). In other words, a hyper-plane is sought which maximizes the soft-margin between the origin and the sample points. Points lying on the "wrong" side of the hyper-plane are outliers. The kernel trick can also be employed for OC-SVM. Schölkopf et. al show that for kernels $k(x, y)$ that depend only on $x - y$, such as the Gaussian kernel, the solutions found by OC-SVM and SVDD are identical. They further showed that the value $\nu = \frac{1}{nC}$, where C is the regularization parameter in the SVM equation, is an upper bound on the number of outliers, a lower bound on the number of support vectors, and that for probability measures P without discrete components, asymptotically the number of outliers and support vectors are equal, in probability. Vert and Vert (2006) correctly point out that while OC-SVM can guarantee the type I error, no guarantees are made regarding consistency of the result (i.e., whether the result converges to a region of minimum volume). This same point is valid for SVDD as well. Indeed, the poor performance of SVDD using polynomial kernels is sufficient proof that the minimum volume set (in the original feature space) is not found. Thus, both of these approaches are implicit, as they do not explicitly solve for the minimum volume set. Similar results for the Minimax Probability Machine (where the type I error is bounded but the resulting set does not necessarily have the minimum volume) are provided by Lanckriet et. al (G. R. G. Lanckriet, Ghaoui, Bhattacharyya, & Jordan, 2002; G. Lanckriet et al., 2002). C. D. Scott and Nowak (2006) overcome these limitations where they use Empirical Risk Minimization to prove consistency (in a distribution free manner) and convergence rates of $\left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ using Structural Risk Minimization for trees (these results aren't distribution free; specifically there is a requirement which can be satisfied if p has no plateaus). C. Scott (2007) expands on this analysis, which served as the basis for the 2-class SVM approach used in (Davenport, Baraniuk, & Scott, 2006), where the second class is the uniform distribution.

The results significantly outperformed those of OC-SVM (i.e. a significantly smaller volume was found for approximately the same type I error).

We now turn our attention to level-set estimation. Let $\mathcal{L}_n(t)$ be the estimation of $\mathcal{L}(t)$ given the n sample points. One of the most common error measures is $\lambda(\mathcal{L}(t) \,\Delta\, \mathcal{L}_n(t))$, where λ is the Lebesgue measure and Δ is the symmetric difference (i.e. $A \,\Delta\, B = (A \backslash B) \bigcup (B \backslash A)$). Another common measure is $H_P(\mathcal{L}(t), t) - H_P(\mathcal{L}_n(t), t)$, where $H_P(S, t) = P(S) - t\lambda(S)$ is the excess mass of S. Both of these measures are non-negative and equal to zero at the optimal solution. Much of the prior work which explicitly solves the level-set estimation problem shows consistency by proving that as n goes to infinity, one of these two measures goes to zero. Most recent work focuses on calculating convergence rates under various conditions on the density p. One of the most common techniques for level-set estimation is the *plug-in estimate* where $\mathcal{L}_n(t) = \{x : \hat{p}_n(x) > t\}$, for a density estimate $\hat{p}_n$ of p. The kernel density estimate (Parzen, 1962) is most often used. For a thorough analysis of the plug-in estimate (in terms of consistency and convergence rates) see Cuevas and Fraiman (1997); Cadre (2006); Rigollet and Vert (2008). Interestingly, the SCC community appears to have been inclined to pursue alternate and novel approaches over the straight-forward use of the kernel density estimate as part of the plug-in estimator. It must be stressed that these approaches have largely been implicit, in the sense that they are based on either a heuristic or some other approximation, and consistency is not proven. For example, Breunig et al. (2000) develop a measure they call the Local Outlier Factor (LOF). LOF is calculated based on a smoothed k-nearest-neighbor distance, where the LOF is calculated as an average ratio of these distances between the neighbors of a point and the point itself. In other words, the LOF is calculated so that objects "deep within a cluster" will have a LOF of approximately 1, while objects near edges of clusters or far from other points will have large values. This seems to be a heuristic way of estimating $f(p)$ where f is hoped to be a monotonically decreasing function. Hempstalk, Frank, and Witten (2008) use the plug-in estimate approach where they use a rather different way of establishing $\hat{p}_n$. Using Minter's notation from above, they generate an artificial distribution for class $\emptyset$, and then it follows from Bayes Theorem that:

$$\Pr\{x|1\} = \frac{\Pr\{\emptyset\} \Pr\{1|x\}}{\Pr\{1\} \Pr\{\emptyset|x\}} \Pr\{x|\emptyset\}.$$

Since the artificial distribution is known, and the prior can be controlled, $\Pr\{x|1\}$ can be estimated from $\Pr\{1|x\}$, which is estimated using class-probability estimation techniques, specifically bagged trees with Laplacian smoothing. In practice, they use a density estimate of p to establish the density for the artificial set. While the technique is certainly interesting, it would be of great interest to see if consistency or convergence rates could be proven. Vert and Vert (2006) demonstrated that one need not estimate the density directly in order

to determine the level-set. They prove that an SVM, with a convex loss function and Gaussian kernel with a "well-calibrated bandwidth σ," can produce an estimate $\mathcal{L}_n(t)$, such that $\lim_{n\to\infty} H_P(\mathcal{L}(t),t) - H_P(\mathcal{L}_n(t),t) = 0$, in probability. Steinwart, Hush, and Scovel (2004) provide convergence rates when using L1-SVM for the error measure $\mu(\mathcal{L}(t) \,\Delta\, \mathcal{L}_n(t))$, where μ is a reference probability distribution.

Finally, we consider other works, starting with transformations to the two-class setting. All of these approaches rely on the creation of a second class in the vicinity of the target class. Examples of this are (Bánhalmi, Kocsor, & Busa-Fekete, 2007) where SVM is used to separate between the two classes, and (Curry & Heywood, 2009), where genetic programming is used and the fitness function accounts for overlap between the two classes. Other works, such as (Rätsch, Mika, Schölkopf, & Müller, 2002), look at how boosting can be applied in the one-class setting. A recent and interesting work is by Juszczak, Tax, Pekalska, and Duin (2009), which uses the premise that the target class should largely be continuous; in other words, if two points belong to the target class, there should be a path from one to the other. For points which are very close to each other, we may expect this to be a straight line. They propose building a minimum spanning tree covering the data, and test membership to the target class by testing the distance of a point to the tree. Since the continuity assumption may be violated for points in different clusters, they allow for the removal of edges in the tree, where longer edges are better candidates for removal. They also allow for a form of dimensionality reduction by removing the shortest paths in the tree. The approach has very good performance on the tested data sets, and it would be of great interest to see if the authors can develop consistency or other theoretical results for it.

4. An Informal Look - an Investment/ROI Analogy

To gain some insight into the one-class classification setting, we now describe an analogous investment game. The learner is given an amount of money to invest, δ. There are N assets which can be invested in, with a cost of p_i to invest in asset i. For each asset i, the learner purchases an amount $r(i) \in [0,1]$ (i.e., from none to all of an asset) and then sells it at a price q_i, determined by the adversary. Any monies not invested are lost. Since the initial wealth is δ, the allocation strategy $r(\cdot)$ must satisfy $\sum_i r(i)p_i \leq \delta$. The overall return to be maximized is $\sum_i r(i)q_i$.

Clearly, the Return-On-Investment (ROI) for asset i is $\frac{q_i}{p_i}$, and thus the learner should invest in assets which have the highest ROI (where free assets are taken to have infinite ROI). In the SCC setting, the fact that the learner must select the investment strategy, $r(\cdot)$, before the adversary determines the selling prices, clearly makes this a difficult proposition. Had we reversed the order, and the adversary were to determine the selling prices first, we

would have a two-class classification problem (i.e., the learner, with full knowledge of both classes, is to minimize type II error subject to a maximum type I error). In this case, the learner's optimal investment strategy would be clear:

> The learner shouldn't invest in an asset k, unless all assets with a higher ROI than k have already been purchased.

Note that while this strategy applies to the soft setting ($r(i) \in [0,1]$), the optimal solution is very nearly identical to that of the hard solution ($r(i) \in \{0,1\}$), with the only difference being that any left over money is invested. How does this investment strategy translate from the two-class classification setting to our original one-class classification setting, where the learner must invest without knowing the ROI values? Clearly, if the adversary's strategy space has some inherent constraints on the relative ROI of assets, then the learner could take advantage of them. For example, in the simplest case, if the adversary's strategy space enforces an ordering on the ROI values, for example $j < k \Rightarrow \frac{q_j}{p_j} < \frac{q_k}{p_k}$, then the learner can invest optimally without knowing Q. However, the less the adversary's strategy space constrains the relative ROI of assets, the more difficult the learner's task is. We would intuitively expect that, in the face of an adversary determined to minimize the learner's return, that less constraints on the adversary would force the learner to diversify his investment. In the extreme case of no constraints at all on the adversary, the learner should purchase the same amount of every non-free asset.[4] We also note that the more the learner diversifies, the "further" his investment strategy becomes relative to the optimal two-class strategy (in accordance to *known* ROI values).

5. On the Optimality of Monotone and Low-Density Rejection Functions

The vast majority of the literature on SCC deals with various techniques for implementing the Low-Density Rejection Strategy (LDRS). This raises the question of whether such a strategy is optimal or not, and under what conditions may it be reasonable to use such a strategy. Since we are interested in adversarial applications, *worst-case* performance is a natural measure for us to consider. For example, if one considers an authentication system every attempt to gain access results in either access being granted or an alarm being fired. From a worst-case perspective, we should expect a sophisticated intruder to be capable of spying upon legitimate use of the system for some period of time and seeing what events or patterns should provide access. Thus, it is more likely that the intruder will attempt to enter a highly probable event in order to gain access, rather than a low-probability event.

4. Note that this is different than 'dollar-cost averaging'; the same amount of money isn't spent on each asset, rather the same absolute amount of each asset is purchased. This guarantees the learner a total ROI of at least 1 (i.e., for every dollar invested, a dollar is earned upon selling).

In fact, the intruder's distribution could be even more concentrated on the highly-probable events than the user's!

Viewed in this perspective, it is not at all clear at the outset that the standard LDRS approach to SCC is the best for adversarial applications. By constraining the adversary's strategy space to one where all of the distributions are tightly concentrated on the highly-probably events under P, low-density-rejection may not be an optimal strategy for the learner. In the extreme case where the adversary always plays the most probable event under P, the adversary would always be able to gain access if the learner plays the low-density-rejection strategy, while potentially the learner could completely deny the adversary access if δ is greater than the probability for that event. Clearly, the nature of the constraints placed on the adversary is critical not only in terms of whether LDRS is optimal, but also in terms of the error that is achievable (both by LDRS and by other strategies). Here we address the former issue, which we feel is of particular relevance considering the large body of existing work which examines approximating low-density rejection functions[5] that can be leveraged in solving practical problems, and leave the latter for future research.

The partially good news is that low-density rejection is worst-case optimal if the learner is confined to "hard" decisions and when the adversary is strong enough in the sense that her strategy space is sufficiently large as shown in Theorem 10. However, as we demonstrate in Section 6, LDRS is inferior in general to the optimal soft strategy. Thus, by playing a randomized strategy, a very significant gain can be achieved.

In this section, we assume a finite support of size N; that is, $\Omega = \{1, \ldots, N\}$ and $P \triangleq \{p_1, \ldots, p_N\}$ and $Q \triangleq \{q_1, \ldots, q_N\}$ are probability mass functions. Note that this assumption still leaves us with an infinite game because the learner's pure strategy space, $\mathcal{R}_\delta(P)$, is infinite. Extensions to infinite support ($N \to \infty$) for many of the finite support results are given in Nisenson (2010). A simple observation is that for any $r \in \mathcal{R}^*_\delta$ there exists $r' \in \mathcal{R}^*_\delta$ such that $r'(i) = r(i)$ for all i such that $p_i > 0$ and for zero probabilities, $p_j = 0$, $r'(j) = 1$. We thus assume w.l.o.g. that $p_i > 0$ for all $i \in \Omega$.

While the low-density rejection strategy implies an assumption that lower probability events should be completely rejected, we instead examine a weaker, but perhaps more useful, condition. Intuitively, it seems plausible that the learner should not assign higher rejection values to higher probability events under P. That is, one may expect that a reasonable rejection function $r(\cdot)$ would be monotonically decreasing with probability values. In the ROI analogy, we would state this as "the learner should prefer cheaper assets to more expensive ones." This is appealing, as more of a cheaper asset can be purchased for the same amount of money than a more expensive asset, and a lower selling price is necessary to achieve the same ROI. We now define two types of monotonicity.

5. See, e.g., (Schölkopf et al., 2001; Cuevas & Fraiman, 1997; Cadre, 2006; Breunig et al., 2000).

Definition 2 (Monotonicity). A rejection function $r(\cdot)$ is *monotone* if $p_j < p_k \Rightarrow r(j) \geq r(k)$. A monotone rejection function $r(\cdot)$ is *strictly monotone* if $p_j = p_k \Rightarrow r(j) = r(k)$.

We note that completely rejecting null-events under P (i.e., $p_j = 0 \Rightarrow r(j) = 1$) does not break strict-monotonicity so our assumption that there are no null events under P is taken w.l.o.g. Surprisingly, optimal monotone strategies are not always guaranteed as shown in the following example.

Example 1 (Non-Monotone Optimality). In the hard setting, take $N = 3$, $P = (0.06, 0.09, 0.85)$ and $\delta = 0.1$. The two δ-valid hard rejection functions are $r' = (1, 0, 0)$ and $r'' = (0, 1, 0)$. Let $\mathcal{Q} = \{Q = (0.01, 0.02, 0.97)\}$. Clearly $\rho(r', Q) = 0.01$ and $\rho(r'', Q) = 0.02$ and therefore, $r''(\cdot)$ is optimal despite breaking monotonicity. More generally, this example holds if $\mathcal{Q} = \{Q : q_2 - q_1 \geq \varepsilon\}$ for any $0 < \varepsilon \leq 1$.

In the soft setting, let $N = 2$, $P = (0.2, 0.8)$, and $\delta = 0.1$. We note that $\mathcal{R}_\delta(P) = \{r^\varepsilon = (0.1 + 4\varepsilon, 0.1 - \varepsilon)\}$, for $\varepsilon \in [-0.025, 0.1]$. We take $\mathcal{Q} = \{Q = (0.1, 0.9)\}$. Then $\rho^\epsilon(Q) = 0.1 + 0.4\varepsilon - 0.9\varepsilon = 0.1 - 0.5\varepsilon$. This is clearly maximized when we minimize ε by taking $\varepsilon = -0.025$, and then the optimal rejection function is $(0, 0.125)$, which clearly breaks monotonicity. This example also holds for $\mathcal{Q} = \{Q : q_2 \geq cq_1\}$ for any $c > 4$.

This example naturally raises the question of which conditions are necessary or sufficient for optimal monotone strategies to be guaranteed. To motivate our sufficient condition for optimality (Property A below), recall the intrusion detection setting discussed in the beginning of this section. There the adversary is constrained to distributions that are tightly concentrated on the highly probably events under P. In this case, since low probability events are scarcely "attacked" by the adversary, the optimal learner would not waste rejection "resources" on low probability events. In other words, in such cases monotone rejection functions aren't optimal. This begets the question if monotone rejection functions are optimal when the adversary is not constrained from attacking low probability events.

Definition 3 (Property A). Let P be a distribution and $\mathcal{Q}$ be a set of distributions. If for all $p_j < p_k$ and $Q \in \mathcal{Q}$ for which $q_j < q_k$, there exists a distribution $Q' \in \mathcal{Q}$ such that for all $i \neq j, k$, $q'_i = q_i$ and $q_j + q'_j \geq q_k + q'_k$, then $\mathcal{Q}$ possesses Property A w.r.t. P.

Example 2 (Possession of Property A). Let P be any distribution over Ω. Let $\mathcal{Q}_1 = \{U\}$, where U is the uniform distribution over Ω. Then $\mathcal{Q}_1$ has Property A w.r.t. P since $q_j < q_k$ is never true. Similarly, let $\mathcal{Q}_2$ be the set of all distributions (if Q is a distribution over Ω, then $Q \in \mathcal{Q}_2$). Then $\mathcal{Q}_2$ also has Property A w.r.t. P. If $P \neq U$, and $\mathcal{Q}_3 = \{P\}$, then, $\mathcal{Q}_3$ doesn't possess Property A w.r.t P.

The following theorem ensures that there exists an optimal monotone rejection function whenever $\mathcal{Q}$ satisfies Property A. In such cases the learner's search space can be conveniently confined to monotone strategies.

Theorem 4 (Optimal Monotone Hard Strategies). When the learner is restricted to hard-decisions and $\mathcal{Q}$ satisfies Property A w.r.t. P, then there exists a monotone $r \in \mathcal{R}^*_\delta$.

Theorem 4 only concerns the hard setting where r is a zero-one rule. The following Property B and the accompanying Theorem 6 treat the more general soft setting.

Definition 5 (Property B). Let P be a distribution and $\mathcal{Q}$ be a set of distributions. If for all $0 < p_j \leq p_k$ and $Q \in \mathcal{Q}$ for which $\frac{q_j}{p_j} < \frac{q_k}{p_k}$, there exists $Q' \in \mathcal{Q}$ such that for all $i \neq j, k$, $q'_i = q_i$ and $\frac{q'_j}{p_j} \geq \frac{q'_k}{p_k}$, then $\mathcal{Q}$ possesses Property B w.r.t. P.

Example 3 (Possession of Property B). Let P be any distribution over Ω. Let $\mathcal{Q}_1 = \{U\}$, $\mathcal{Q}_2$ be the set of all distributions and $\mathcal{Q}_3 = \{P\}$. All three sets, $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$, have Property B w.r.t. P.

Recalling our informal investment analogy, if the strategy space of the adversary satisfies Property B, then cheaper assets always have the potential for higher ROI (and equally priced assets have equal ROI opportunities). If this is the case, then Theorem 6 states that there is an optimal investment strategy (that maximizes the overall return), which never purchases more of an expensive asset than a cheaper one and always invests identically in equally priced assets.

Theorem 6 (Optimal Monotone Soft Strategies).
If $\mathcal{Q}$ satisfies Property B w.r.t. P, then there exists an optimal strictly monotone rejection function.

Remark 7. It is not hard to prove that a slightly stronger version of Property A implies Property B. The stronger version of Property A is that the property also holds when $p_j = p_k$ (rather than only for $p_j < p_k$).

In the remainder of this section we only consider the hard setting. Theorem 4 tells us that there exists an optimal rejection function in the set of monotone rejection functions provided that Property A holds. Obviously, to be optimal the rejection function should reject as much as possible up to the δ bound. We now show that if $\mathcal{Q}$ is sufficiently rich (satisfying Property C below) then any "low-density rejection function" is optimal.

Definition 8 (Low-Density Rejection Function (LDRF) and Strategy (LDRS)). A hard, δ-valid, monotone rejection function $r(\cdot)$ is called a *low-density rejection function* if its $\rho(r, P)$ is maximal among all hard, monotone δ-valid rejection functions. The strategy of selecting any LDRF is called the *low-density rejection strategy (LDRS).*

Definition 9 (Property C). Let P be a distribution. We say that the set $\mathcal{Q}$ satisfies Property C (w.r.t. P) if for each $p_j = p_k$ and $Q \in \mathcal{Q}$, there exists $Q' \in \mathcal{Q}$ such that $q'_j = q_k$ and $q'_k = q_j$, and for all other events, Q' identifies with Q.

Some intuition about Property C can be gained by considering some adversary strategy space $\mathcal{Q}$. First note that by expanding $\mathcal{Q}$ to satisfy Property C the adversary can only be strengthened. The property ensures that the adversary can take advantage of situations where the learner doesn't identically treat equally probable events under P. When the adversary is sufficiently strong in this sense we are able to show that LDRS dominates any monotone rejection function. Therefore, if $\mathcal{Q}$ also satisfies Property A, in which case there exists an optimal monotone rejection function (Theorem 4), then LDRS is optimal. This is summarized in the following theorem.

Theorem 10 (LDRS Optimality). Let r^* be an LDRF. Let r be any monotone δ-valid rejection function. Then, r^* dominates r,

$$\min_{Q\in\mathcal{Q}} \rho(r^*, Q) \geq \min_{Q\in\mathcal{Q}} \rho(r, Q), \tag{1}$$

for any $\mathcal{Q}$ satisfying Property C. Thus, if $\mathcal{Q}$ possess both Property A and Property C w.r.t. P, then LDRS is hard-optimal.

Example 4 (Violating Property C Breaks Domination). We illustrate here a violation of Property C may result in a violation of the domination inequality (1) in Theorem 10. Let $N = 5$, $P = (0.02, 0.03, 0.05, 0.05, 0.85)$, and $\delta = 0.1$. Then the two δ-valid LDRS rejection functions are $r = (1, 1, 1, 0, 0)$ and $r' = (1, 1, 0, 1, 0)$. Let $\mathcal{Q} = \{Q : q_3 - q_4 > \varepsilon\}$ for some $0 < \varepsilon < 1$. Clearly, $\mathcal{Q}$ does not satisfy Property C. For any $Q \in \mathcal{Q}$, $\rho(r, Q) - \rho(r', Q) = q_3 - q_4 > \varepsilon$, and therefore, $\min_{Q\in\mathcal{Q}} \rho(r', Q) < \min_{Q\in\mathcal{Q}} \rho(r, Q)$. Thus, the monotone function r dominates the LDRF, r'. Hence, LDRS isn't optimal because r' could be chosen.

6. The Omniscient Adversary: Games, Strategies and Bounds

We next turn our attention to the power of the adversary, an issue that hasn't been emphasized in the SCC literature, but has crucial impact on the relevancy of SCC solutions in adversarial applications. For example, when considering intrusion detection (see, e.g., Lazarevic, Ertöz, Kumar, Ozgur, & Srivastava, 2003), it is necessary to assume that the "attacking distribution" has some worst-case characteristics and it is important to quantify precisely what the adversary knows or can do. The simple observation in this setting is that an *omniscient and unconstrained adversary*, who knows all parameters of the game including the learner's strategy, would completely demolish the learner who uses hard strategies. By using a soft strategy, the learner can achieve the slightly better result of $1 - \delta$ type II error (false negative rate). In either case, the presence of such a powerful adversary makes the SCC problem trivial and the resulting rejection function is practically worthless. These simple results are developed in Section 6.1.

We therefore consider an omniscient but limited adversary. In seeking a useful and quantifiable constraint on $\mathcal{Q}$ it is helpful to recall that the essence of the SCC problem is to

try to distinguish between two probability distributions (albeit one of them unknown). A natural constraint is a lower bound on the "distance" between these distributions. Indeed, it is immediately obvious that if $P \in \mathcal{Q}$, the adversary can always achieve the maximal type II error of $1-\delta$ by selecting $Q = P$. Following similar results in hypothesis testing (see Cover & Thomas, 1991, Chapt. 12), we could consider games in which the adversary must select Q such that $D(P||Q) \geq \Lambda$, for some constant $\Lambda > 0$, where $D(\cdot||\cdot)$ is the KL-divergence; that is, $D(P||Q) \triangleq \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$ (Cover & Thomas, 1991). Unfortunately, this constraint is vacuous since $D(P||Q)$ "explodes" when $q_i \ll p_i$ (for any i). In this case the adversary can optimally play the same strategy as in the unrestricted game while meeting the KL-divergence constraint. Fortunately, by taking $D(Q||P) \geq \Lambda$, we can effectively constrain the adversary.[6] Instead of only considering the KL-divergence we consider adversary constraints using a large family of divergences that include the KL-divergence, the L_2 norm and various Bregman divergences. Definitions 11 and 13 characterize this family.

One of our main contributions is a complete analysis of this constrained game in Section 6.2, including identification of the optimal strategy for the learner and the adversary, as well as the best achievable false negative rate. The optimal learner strategy and best achievable rate are obtained via a solution of a linear program specified in terms of the problem parameters. These results are immediately applicable as *lower bounds* for standard (finite-sample) SCC problems, but may also be used to inspire new types of algorithms for standard SCC. While we do not have a closed form expression for the best achievable false-negative rate, we provide a few numerical examples demonstrating and comparing the optimal "hard" and "soft" performance.

6.1 Unrestricted Adversary

In the first game we analyze an adversary who is completely unrestricted. This means that $\mathcal{Q}$ is the set of all distributions. Unsurprisingly, this game leaves little opportunity for the learner. For any rejection function $r(\cdot)$, define $r_{min} \triangleq \min_i r(i)$ and $I_{min}(r) \triangleq \{i : r(i) = r_{min}\}$. For any distribution D, $\rho(r, D) = \sum_{i=1}^{N} d_i r(i) \geq \sum_{i=1}^{N} d_i r_{min} = r_{min}$, in particular, $\delta = \rho(r, P) \geq r_{min}$ and $\min_Q \rho(r, Q) \geq r_{min}$. By choosing Q such that $q_i = 1$ for some $i \in I_{min}(r)$, the adversary can achieve $\rho(r, Q) = r_{min}$ (the same rejection rate is achieved by taking any Q with $q_i = 0$ for all $i \notin I_{min}(r)$). In the soft setting, $\min_Q \rho(r, Q)$ is maximized by the rejection function $r^{\delta}(i) \triangleq \delta$ for all $p_i > 0$ ($r^{\delta}(i) \triangleq 1$ for all $p_i = 0$). This is equivalent to flipping a δ-biased coin for non-null events (under P). The best achievable type II error is

6. Under the investment analogy, requiring that $D(P||Q)$ be large is equivalent to requiring a small "average" value for $\frac{q_i}{p_i}$ (giving the learner poor investment opportunities). On the other hand, requiring that $D(Q||P)$ be large is equivalent to requiring that the "average" value of $\frac{q_i}{p_i}$ be sufficiently large (providing the learner with good investment opportunities, and potentially increasing the value of $\rho(r, Q)$).

$1-\delta$. In the hard setting, clearly $r_{min}=0$ (otherwise $1>\delta\geq 1$), and the best achievable type II error is precisely 1. That is, absolutely nothing can be achieved.

This simple analysis shows the futility of the SCC game when the adversary is too powerful. In order to consider SCC problems at all one must consider reasonable restrictions on the adversary that lead to more useful games. One type of such a restriction would be to limit the adversary's knowledge of $r(\cdot)$, P and/or of δ. Another type would be to directly limit the strategic choices available to the adversary. We note that the former type of restriction doesn't affect the best achievable type II error, and thus in the next section we will focus on the latter.

6.2 An Omniscient, but Constrained, Adversary

While we could therefore define $\mathcal{Q}=\mathcal{Q}_\Lambda\triangleq\{Q:D(Q||P)\geq\Lambda\}$, we instead will consider a more general family. First, let $\mathcal{X}$ be the N-dimensional simplex: $\mathcal{X}\triangleq\{(x_1,\ldots,x_N):x_i\geq 0,\sum_{i=1}^N x_i=1\}$. For convenience, we now define a transfer function, $t(X,a,b)\to\mathcal{X}$, where $X\in\mathcal{X}$, and a and b are indices in $\{1,\ldots,N\}$, which transfers probability from event b to event a, as:

$$t(X,a,b)_i=\begin{cases}x_a+x_b & i=a,\\ 0 & i=b,\\ x_i & \text{otherwise.}\end{cases}$$

Definition 11 (2-Symmetric). A function $f_P:\mathcal{X}\to\mathbb{R}$, is called *2-symmetric* if for all $X\in\mathcal{X}$ and for all j,k such that $p_j=p_k$, $f_P\left(t(X,j,k)\right)=f_P\left(t(X,k,j)\right)$.

Remark 12. We note that a Bregman divergence (defined over $[0,1]^N$) may be 2-symmetric. Specifically, define $D_P(Q)=B_F(Q||P)\triangleq F(Q)-F(P)-\nabla F(P)\cdot(Q-P)$. Let $\Delta_F\triangleq F(t(Q,j,k))-F(t(Q,k,j))$. Then, the divergence is 2-symmetric if:

$$\begin{aligned}0=D_P(t(Q,j,k))-D_P(t(Q,k,j))=&\Delta_F-\nabla F(P)\cdot(t(Q,j,k)-t(Q,k,j))\\ =&\Delta_F+(q_j+q_k)\left(\frac{\partial F(P)}{\partial x_k}-\frac{\partial F(P)}{\partial x_j}\right).\end{aligned}$$

We note that if $F(X)=\sum_{i=1}^N f(x_i)$, where $f(\cdot)$ is a strictly convex function, then clearly the Bregman divergence is 2-symmetric.

Definition 13 (Receding). A function $f_P:\mathcal{X}\to\mathbb{R}$, is called *receding* if for all $X\in\mathcal{X}$, $p_j<p_k$ and $x_k>0$, $f_P(t(X,j,k))>f_P(X)$. A receding function $D_P:\mathcal{X}\to\mathbb{R}$ is called a *receding divergence* if it is defined over the domain $[0,1]^N$, it is differentiable over $(0,1)^N$ and is strictly convex.

Remark 14. We note that a Bregman divergence may be a receding divergence, as well. Specifically, define $D_P(Q) = B_F(Q||P) \triangleq F(Q) - F(P) - \nabla F(P) \cdot (Q - P)$. This trivially meets the differentiability and strict convexity requirements. Let us examine if it is receding. Let $p_j < p_k$, $q_k > 0$ and let $\Delta \triangleq t(Q, j, k) - Q$. Then, in order to satisfy the property:

$$\begin{aligned} 0 < D_P(t(Q,j,k)) - D_P(Q) =& F(Q+\Delta) - F(Q) - \nabla F(P) \cdot \Delta \\ =& F(Q+\Delta) - F(Q) + q_k \left(\frac{\partial F(P)}{\partial x_k} - \frac{\partial F(P)}{\partial x_j} \right). \end{aligned}$$

We note that if $F(X) = \sum_{i=1}^{N} f(x_i)$, where $f(\cdot)$ is a strictly convex function, then $F(t(X,j,k)) = F(t(X,k,j))$ for all j, k, and thus, by convexity:

$$\begin{aligned} F(Q+\Delta) - F(Q) = F(t(Q,j,k)) - F(Q) \geq 0 \\ \frac{\partial F(P)}{\partial x_k} - \frac{\partial F(P)}{\partial x_j} = f'(p_k) - f'(p_j) > 0. \end{aligned}$$

Thus, Bregman divergences which are of this form, such as the squared Euclidean distance $D_P(Q) = ||Q - P||^2$ and the KL-Divergence, are also (2-symmetric) receding divergences. Note that this condition is sufficient and not necessary. It is certainly possible for Bregman divergences which are not of this form to be receding divergences as well.

We define $\mathcal{Q}_\Lambda \triangleq \{Q : D_P(Q) \geq \Lambda\}$, where $D_P(\cdot)$ is a 2-symmetric receding divergence. We say that a distribution Q *meets the divergence constraint* if $D_P(Q) \geq \Lambda$. As we will shortly see, this is consistent with an adversary that can't eavesdrop on the user, as the constraint prevents the adversary from selecting distributions which are only concentrated on high-probability events under P.

Lemma 15. $\mathcal{Q}_\Lambda$ possesses Properties A and B w.r.t. P.

Proof Let j, k be such that $p_j \leq p_k$. For any distribution $Q \in \mathcal{Q}_\Lambda$ we define $Q' = t(Q, j, k)$. If $p_j < p_k$, then since $D_P(\cdot)$ is receding, $D_P(Q') \geq D_P(Q) \geq \Lambda$. Otherwise, if $p_j = p_k$, since $D_P(\cdot)$ is 2-symmetric and convex, $D_P(Q') \geq D_P(Q) \geq \Lambda$. Thus, in either case, $Q' \in \mathcal{Q}_\Lambda$. If Q is such that $q_j < q_k$, then $q'_j + q_j = 2q_j + q_k \geq q_k = q'_k + q_k$, and $\mathcal{Q}_\Lambda$ has Property A. If Q is such that $\frac{q_j}{p_j} < \frac{q_k}{p_k}$, then $\frac{q'_j}{p_j} = \frac{q_j + q_k}{p_j} \geq 0 = \frac{q'_k}{p_k}$ and $\mathcal{Q}_\Lambda$ possesses Property B. ■

Therefore, by Theorems 4 and 6 there exists a (strictly) monotone $r \in \mathcal{R}^*_\delta$ in the hard (respectively, soft) setting. If Q_Λ has Property C as well, then by Theorem 10 any δ-valid LDRF is hard-optimal. It is easy to verify that Bregman divergences of the form described in Remark 14 possess Property C.

We now define $X^{(j)}$ as the distribution which is completely concentrated on event j. In other words $x_i^{(j)} \triangleq \mathbb{I}(i = j)$, where $\mathbb{I}(\cdot)$ is the indicator function. We assume that $0 < p_1 \leq p_2 \leq \cdots \leq p_N$. Therefore, since $D_P(\cdot)$ is receding, $D_P\left(X^{(1)}\right) \geq D_P\left(X^{(2)}\right) \geq \cdots \geq$

$D_P\left(X^{(N)}\right)$. Therefore if $D_P\left(X^{(N)}\right) \geq \Lambda$, then any Q that is concentrated on a single event meets the constraint $D_P(Q) \geq \Lambda$. Then, the adversary can play the same strategy as in the unrestricted game, and the learner should select r^δ as before. For the game to be non-trivial it is thus required that $\Lambda > D_P\left(X^{(N)}\right)$. Similarly, if the optimal r is such that there exists $j \in I_{min}(r)$ (that is $r(j) = r_{min}$) and $D_P\left(X^{(j)}\right) \geq \Lambda$, then a distribution Q that is completely concentrated on j has $D_P(Q) \geq \Lambda$ and achieves $\rho(r, Q) = r_{min}$, as in the unrestricted game. Therefore, $r = r^\delta$, and so maximizes r_{min}. This yields the following definition:

Definition 16. A rejection function r is called *vulnerable* if there exists $j \in I_{min}(r)$ such that $D_P\left(X^{(j)}\right) \geq \Lambda$.

We begin our analysis of the game by identifying some useful characteristics of optimal adversary strategies under the assumption that the chosen rejection function isn't vulnerable. These properties, that are stated in Lemma 18, are then used to prove Theorem 19 showing that the effective support of an optimal Q has a size of two at most. Based on these properties, we provide in Theorem 23 a linear program that computes an optimal rejection function (under the assumption that it isn't vulnerable).Finally, in Lemma 24 we show that the solution computed by the linear program is r^δ if it is vulnerable, giving optimal (though trivial) performance. Thus, in any case, the output of the linear program is optimal.

If $\Lambda > D_P\left(X^{(1)}\right)$, then no adversary distribution can meet the divergence constraint. We therefore limit ourselves to cases where $\Lambda \leq D_P\left(X^{(1)}\right)$. We can now divide the events in Ω into two groups: H and L, such that $H = \{i : D_P\left(X^{(i)}\right) < \Lambda\}$ and $L = \Omega \setminus H$. We note that the assumption that r isn't vulnerable implies that $I_{min}(r) \subseteq H$. By definition, $\forall h \in H, l \in L$, we have that $p_h > p_l$.

Lemma 17. If Q meets the divergence constraint, there exists an event $i \in L$ for which $q_i > 0$.

Proof Let us assume that $q_i = 0$ for all $i \in L$. Let j be the smallest event in H. Since $D_P(\cdot)$ is receding, $D_P(Q) \leq D_P\left(X^{(j)}\right) < \Lambda$. Contradiction. ■

Lemma 18. Let r be a rejection function which isn't vulnerable. If Q meets the divergence constraint and minimizes $\rho(r, Q')$:

i. $D_P(Q) = \Lambda$;

ii. Let u, v be two indices in $\{1, \ldots, N\}$. Define $Q'' = t(Q, u, v)$. If $q_v > 0$ and $D_P(Q'') \geq \Lambda$, then $r(u) \geq r(v)$. Furthermore, $r(u) = r(v) \Rightarrow D_P(Q'') = \Lambda$;

iii. $p_j < p_k$ and $q_k > 0 \Rightarrow r(j) > r(k)$;

iv. $p_j < p_k$ and $q_j, q_k > 0 \Rightarrow \frac{\partial D_P(Q)}{\partial x_j} > \frac{\partial D_P(Q)}{\partial x_k}$;

v. $q_j, q_k > 0 \Rightarrow p_j \neq p_k$;

vi. $p_j < p_k$ and $q_j > 0 \Rightarrow D_P(Q) > D_P(t(Q,k,j))$.

Proof

i. Assume that $D_P(Q) > \Lambda$. By Lemma 17 there exists a non-empty set $L_Q \triangleq \{i \in L \mid q_i > 0\}$. Let $h_{max} = \text{argmax}_{i \in I_{min}(r)} q_i$. Clearly, $h_{max} \in H$. We define a new distribution Q^*, which is identical to Q except that probability is transferred from events in L_Q to h_{max}, in order to make $D_P(Q^*) = \Lambda$ (this is possible, since $D_P(\cdot)$ is continuous and, by Lemma 17, transferring all probability from L_Q to h_{max} would result in $D_P(\cdot) < \Lambda$). Since transferring any probability from $i \in L_Q$ to h_{max} results in making $\rho(r, Q)$ smaller, $\rho(r, Q^*) < \rho(r, Q)$, contradicting the fact that Q minimizes $\rho(r, Q')$.

ii. We note that $\rho(r, Q'') = \rho(r, Q) - q_v(r(v) - r(u))$. Since $\rho(r, Q)$ is minimal and $D_P(Q'') \geq \Lambda$ it follows that $r(u) \geq r(v)$. If $r(u) = r(v)$ then $\rho(r, Q'') = \rho(r, Q)$, and by part (i), $D_P(Q'') = \Lambda$.

iii. By part (ii), taking $u = j$ and $v = k$ we trivially get $r(j) \geq r(k)$. Furthermore, since $p_u = p_j < p_k = p_v \Rightarrow D_P(Q'') > \Lambda$, $r(j) \neq r(k)$. Thus, $r(j) > r(k)$.

iv. Assume, contradictorily, that $\frac{\partial D_P(Q)}{\partial x_j} \leq \frac{\partial D_P(Q)}{\partial x_k}$. Let $0 < \epsilon \leq \min\{q_j, q_k\}$. We define $\epsilon_{j,k} = \epsilon \left(X^{(k)} - X^{(j)}\right)$. Then, by convexity:

$$\begin{aligned} D_P(Q + \epsilon_{j,k}) &\geq D_P(Q) + \nabla D_P(Q) \cdot \epsilon_{j,k} \\ &= D_P(Q) + \epsilon \left(\frac{\partial D_P(Q)}{\partial x_k} - \frac{\partial D_P(Q)}{\partial x_j} \right) \\ &\geq D_P(Q). \end{aligned}$$

Therefore, by defining $Q' = Q + \epsilon_{j,k}$, we have that $D_P(Q') \geq D_P(Q) \geq \Lambda$. Furthermore, by part (iii), $r(j) > r(k)$. Therefore, $\rho(r, Q') = \rho(r, Q) + \epsilon(r(k) - r(j)) < \rho(r, Q)$. Contradiction.

v. Assume that $p_j = p_k$. We consider two cases. In the first case, $r(j) < r(k)$, w.l.o.g. By defining $u = j$, $v = k$, from part (ii) we get that $r(j) \geq r(k)$, which is a contradiction. In the second case, $r(j) = r(k)$. However, since both q_j and q_k are greater than zero, defining $u = j$ and $v = k$ in part (ii) gives us that $D_P(Q'') > \Lambda$, which is again a contradiction.

vi. If $q_k = 0$ then $Q = t(t(Q,k,j),j,k)$ and $D_P(Q) > D_P(t(Q,k,j))$. Otherwise, $q_k > 0$ and by part (iii), $r(j) > r(k)$. If we assume contradictorily that $D_P(t(Q,k,j)) \geq D_P(Q) = \Lambda$, then by part (ii), taking $u = k$ and $v = j$, $r(k) \geq r(j)$. Contradiction.

■

Theorem 19. If r isn't vulnerable, then any optimal adversarial strategy Q has an effective support of size at most two.

Proof Let us assume, by contradiction, that the theorem's statement is wrong; that is, there exists an optimal Q^* that has $J > 2$ events for which $q_i^* \neq 0$. W.l.o.g. we rename our events such that these are the first J events. We note that Q^* is a solution (i.e., global minimum) to the following problem $(*)$:

$$\begin{aligned} \text{minimize } \rho(r,Q) = \sum_{i=1}^{J} r(i)q_i, \text{ subject to:} \\ \sum_{i=1}^{J} q_i = 1,\ D_P(Q) = \Lambda, \\ 0 < q_i < 1,\ i \in \{1,\ldots,J\}. \end{aligned}$$

We will now prove that Q^* does not in fact solve the problem. We do so in two parts:

1. We show that Q^* is the unique global maximum of the Lagrangian of $(*)$.
2. We show that there exists a different distribution $\tilde{Q}$ with the same effective support, which meets the equality constraints. We therefore conclude that $\rho(r,\tilde{Q}) < \rho(r,Q)$, contradicting the optimality of Q^*.

We now prove the first part. The Jacobian matrix for the equality constraints at Q^* is:

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ \frac{\partial D_P(Q^*)}{\partial x_1} & \frac{\partial D_P(Q^*)}{\partial x_2} & \frac{\partial D_P(Q^*)}{\partial x_3} & \ldots & \frac{\partial D_P(Q^*)}{\partial x_J} \end{pmatrix}.$$

Since all $q_i^* > 0$, by parts (v) and (iv) of Lemma 18, for all $j,k \leq J$: $p_j \neq p_k$ and $\frac{\partial D_P(Q^*)}{\partial x_j} \neq \frac{\partial D_P(Q^*)}{\partial x_k}$. Therefore, the gradients of the constraints are linearly independent at Q^* and therefore, since Q^* is (at least) a local minimum to the problem $(*)$, there exists a unique Lagrangian multiplier vector $\lambda = (\lambda_1, \lambda_2)$ such that $Q^* = (q_1^*, q_2^*, \ldots, q_J^*)$ is an extremum point of the Lagrangian:

$$L(Q,\lambda) = \sum_{i=1}^{J} r(i)q_i + \lambda_1 \left(D_P(Q) - \Lambda\right) + \lambda_2 \left(\sum_{i=1}^{J} q_i - 1\right).$$

The partial derivatives are: $\frac{\partial L(Q^*,\lambda)}{\partial q_i} = r(i) + \lambda_1 \frac{\partial D_P(Q^*)}{\partial x_i} + \lambda_2 = 0$. Therefore, for all $j,k \in \{1,\ldots,J\}$:

$$\begin{aligned} r(j) + \frac{\partial D_P(Q^*)}{\partial x_j} + \lambda_2 &= r(k) + \lambda_1 \frac{\partial D_P(Q^*)}{\partial x_k} + \lambda_2 \\ \Rightarrow\ \lambda_1 \left(\frac{\partial D_P(Q^*)}{\partial x_j} - \frac{\partial D_P(Q^*)}{\partial x_k} \right) &= r(k) - r(j) \\ \Rightarrow\ \lambda_1 &= \frac{r(k) - r(j)}{\frac{\partial D_P(Q^*)}{\partial x_j} - \frac{\partial D_P(Q^*)}{\partial x_k}} \end{aligned}$$

If we assume (w.l.o.g.) that $p_k < p_j$, then, from parts (iii) and (iv) of Lemma 18, $r(k) > r(j)$ and $\frac{\partial D_P(Q^*)}{\partial x_k} > \frac{\partial D_P(Q^*)}{\partial x_j}$. Thus, $\lambda_1 < 0$. Therefore, due to the strict convexity of $D_P(\cdot)$ and the linearity of the other two equations, the Lagrangian $L(Q,\lambda)$ is strictly concave. Therefore, since Q^* is an extremum point of the (strictly concave) Lagrangian function, it is the unique global maximum.

We now wish to show that there exists some other distribution $\tilde{Q}$ that meets the divergence constraint and has the same support as Q^*. We define Q^{123} as $q_i^{123} = \mathbb{I}(i>3)q_i^*$ and $c_{123} \triangleq q_1^* + q_2^* + q_3^*$. Then we define:

$$\begin{aligned} g(q_1,q_2) &\triangleq Q^{123} + q_1 X^{(1)} + q_2 X^{(2)} + (c_{123} - q_1 - q_2) X^{(3)} \\ f(q_1,q_2) &\triangleq D_P\left(g(q_1,q_2)\right) - \Lambda \\ \Rightarrow \text{for } i \in \{1,2\}:\ \frac{\partial f(q_1,q_2)}{\partial q_i} &= \nabla D_P\left(g(q_1,q_2)\right) \cdot \left(X^{(i)} - X^{(3)}\right) \\ &= \frac{\partial D_P(g(q_1,q_2))}{\partial x_i} - \frac{\partial D_P(g(q_1,q_2))}{\partial x_3} \end{aligned}$$

Clearly, $g(q_1^*,q_2^*) = Q^*$ and $f(q_1^*,q_2^*) = 0$. From part (iv) of Lemma 18, we have for $i \in \{1,2\}$:

$$\frac{\partial f(q_1^*,q_2^*)}{\partial q_i} = \frac{\partial D_P(Q^*)}{\partial x_i} - \frac{\partial D_P(Q^*)}{\partial x_3} \neq 0.$$

Therefore, f is smooth in the open, convex domain $\{q_1,q_2>0\} \cap \{q_1+q_2 < c_{123}\}$ and has a root in this domain at (q_1^*,q_2^*) at which none of its partial derivatives are 0. Then, there exist an infinite number of points in the domain for which $f=0$ (this is true for any sub-domain for which (q_1^*,q_2^*) is an interior point). Let $(\tilde{q}_1,\tilde{q}_2) \neq (q_1^*,q_2^*)$ be one of these points. Then, the distribution $\tilde{Q} = (\tilde{q}_1,\tilde{q}_2,c_{123}-\tilde{q}_1-\tilde{q}_2,q_4^*,q_5^*,\ldots,q_J^*) \neq Q^*$ satisfies $D(\tilde{Q},P) = \Lambda$ and has the exact same effective support as Q^*. Therefore, $\tilde{Q}$ meets the equality criteria of the Lagrangian. Since Q^* is the unique global maximum of $L(Q,\lambda)$: $\rho(r,\tilde{Q}) = L(\tilde{Q},\lambda) < L(Q^*,\lambda) = \rho(r,Q^*)$, contradicting the fact that Q^* is optimal. ■

We now turn our attention to the learner's selection of $r(\cdot)$. As already established by Lemma 15 and Theorem 6, it is sufficient for the learner to consider only strictly monotone

rejection functions. Since for these functions $p_j = p_k \Rightarrow r(j) = r(k)$, the learner can partition Ω into $K = K(P, \Omega)$ event subsets, which correspond, by probability, to "level sets", $S_1, S_2, \ldots, S_K$ (all events in a level set S_j have probability $p^{(S_j)}$). We re-index these subsets such that $0 < p^{(S_1)} < p^{(S_2)} < \cdots < p^{(S_K)}$. Define K variables $r_1, r_2, \ldots, r_K$, representing the rejection rate assigned to each of the K level sets ($\forall \omega \in S_i, r(\omega) = r_i$). Since $D_P(\cdot)$ is 2-symmetric, $D_P\left(X^{(\omega)}\right)$ is constant for all ω in a level set S. Therefore, we use the notation $D_P^S \triangleq D_P\left(X^{(\omega)}\right)$ for any $\omega \in S$. We group our level sets by probability: $\mathcal{L} = \{S : D_P^S > \Lambda\}$, $\mathcal{M} = \{S : D_P^S = \Lambda\}$, and $\mathcal{H} = \{S : D_P^S < \Lambda\}$. We define $w \triangleq \text{argmax}_i\{S_i \in \mathcal{L} \bigcup \mathcal{M}\}$.

Lemma 20. If Q minimizes $\rho(r, Q)$ and meets the constraint $D_P(Q) \geq \Lambda$, then $r_w \geq \rho(r, Q)$.

Proof Let $j \in S_w$. Then $D_P\left(X^{(j)}\right) \geq \Lambda$, and since Q minimizes $\rho(r, Q)$, $r_w = \rho\left(r, X^{(j)}\right) \geq \rho(r, Q)$. ■

By Theorem 19, if r isn't vulnerable, the adversary-optimal Q will have an effective support of at most size 2. If it has an effective support of size 1, then the event ω for which $q_\omega = 1$ cannot be from a level set in $\mathcal{L}$ or $\mathcal{H}$ (otherwise, part (i) of Lemma 18 would be violated). Therefore, it must belong to the single level set in $\mathcal{M}$. Thus, if $\mathcal{M} = \{S_m\}$ (for some index m), there are feasible solutions Q such that $q_\omega = 1$ (for $\omega \in S_m$), all of which have $\rho(r, Q) = r_m$. The following lemma characterizes optimal distributions Q which have an effective support of size 2.

Lemma 21. If r isn't vulnerable and Q is optimal with an effective support of size 2 (that is, there are j, k such that $q_j, q_k > 0$ and $q_j + q_k = 1$), then, assuming w.l.o.g. that $p_j \leq p_k$, $j \in S_l \in \mathcal{L}$ and $k \in S_h \in \mathcal{H}$ for some l and h.

Proof Since $q_j, q_k > 0$, and Q is optimal, we have that $p_j \neq p_k$, by part (v) of Lemma 18. Therefore, $p_j < p_k$, and by part (vi) of Lemma 18,

$$D_P\left(X^{(k)}\right) = D_P(t(Q, k, j)) < D_P(Q) < D_P(t(Q, j, k)) = D_P\left(X^{(j)}\right).$$

Assume, by contradiction, that k belongs to a level set in $\mathcal{L}$ or $\mathcal{M}$. This is equivalent to $D_P\left(X^{(k)}\right) \geq \Lambda$. We therefore have that $D_P(Q) > D_P\left(X^{(k)}\right) \geq \Lambda$, which is a violation of part (i) of Lemma 18. Therefore, k belongs to a level set in $\mathcal{H}$. Likewise, were we to assume that j belongs to a level set in $\mathcal{M}$ or $\mathcal{H}$ ($D_P\left(X^{(j)}\right) \leq \Lambda$), it would follow that $D_P(Q) < D_P\left(X^{(j)}\right) \leq \Lambda$, which would also violate part (i) of Lemma 18. Therefore, j belongs to a level set in $\mathcal{L}$. ■

Lemma 22. Let $S_l \in \mathcal{L}$ and $S_h \in \mathcal{H}$. Then, there always exists a single solution $q_\Lambda^{(l,h)} \in (0,1)$ to

$$D_P\left(qX^{(j)} + (1-q)X^{(k)}\right) = \Lambda,$$

for any $j \in S_l, k \in S_h$.

Proof Let Q be a distribution with an effective support of size 2, where the events j,k for which $q_j, q_k > 0$ are such that $j \in S_l$ and $k \in S_h$. Furthermore, let $q_j = q$ and $q_k = 1-q$. Define $g(q) \triangleq g(q,j,k) \triangleq D_P\left(qX^{(j)} + (1-q)X^{(k)}\right)$. Then, $g(q) = D_P(Q)$. We note that $g(0) = D_P^{S_h} < \Lambda$ and $g(1) = D_P^{S_l} > \Lambda$. Thus a solution, q^*, exists in the range $(0,1)$. Since $g(q)$ is continuous and convex, there cannot exist another solution in this range. Let $X = q^*X^{(j)} + (1-q^*)X^{(k)}$. Let $j' \in S_l$ and $k' \in S_h$. Then, since $D_P(\cdot)$ is 2-symmetric, $\Lambda = D_P(X) = D_P(t(X,j',j)) = D_P(t(X,k',k)) = D_P(t(t(X,j',j),k',k))$, and thus the solution is the same for all pairs of members between S_l and S_h. ■

Therefore, if an adversary-optimal Q has an effective support of size 2, where the events with non-zero probability are from S_l and S_h respectively, then, $\rho(r,Q) = \rho^{(l,h)} \triangleq q_\Lambda^{(l,h)} r_l + (1 - q_\Lambda^{(l,h)}) r_h$.

Therefore, the adversary's choice of an optimal distribution, Q, must have one of $|\mathcal{L}||\mathcal{H}| + |\mathcal{M}| \leq \lfloor \frac{K^2}{4} \rfloor$ (possibly different) rejection rates. Each of these rates, $\rho_1, \rho_2, \ldots, \rho_{|\mathcal{L}||\mathcal{H}|+|\mathcal{M}|}$, is a linear combination of at most two variables, r_i and r_j. We introduce an additional variable, z, to represent the max-min rejection rate. This entails the following theorem.

Theorem 23. An optimal soft rejection function and the lower-bound on the optimal type II error, $1-z$, is obtained by solving the following linear program:

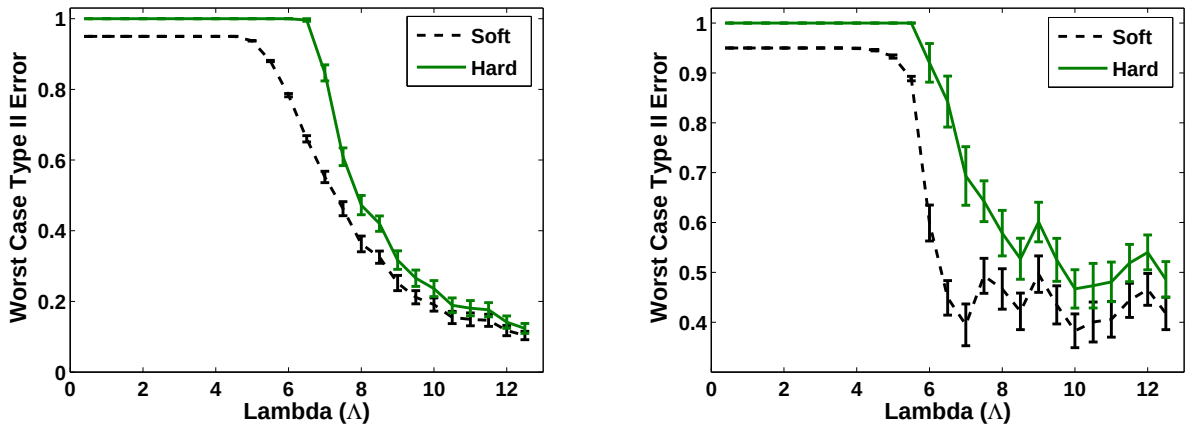
$$\begin{aligned} \text{maximize}_{r_1,r_2,\ldots,r_K,z} \;\; & z, \text{ subject to:} \\ & \sum_{i=1}^{K} r_i |S_i| p(S_i) = \delta \\ & 1 \geq r_1 \geq r_2 \geq \cdots \geq r_K \geq 0 \\ & r_w \geq z \\ & \rho_i \geq z, \; i \in \{1,2,\ldots,|\mathcal{L}||\mathcal{H}| + |\mathcal{M}|\}. \end{aligned} \tag{2}$$

Let r^* be the solution to the linear program (2). Our derivation of the linear program is dependent on the restriction that r^* isn't vulnerable. If r^* contradicts this restriction then, as discussed, the optimal strategy is r^δ. The following lemma shows that in this case $r^* = r^\delta$ anyway, and thus the solution to the linear program is always optimal. Its proof can be found in Appendix B.

Lemma 24. Let r^* be the solution to the linear program. If r^* is vulnerable, then $r^* = r^\delta$.

Remark 25. We attempted to determine explicit bounds on the value of $1-z$, the optimal type II error, that would result from solving the linear program in Theorem 23, including via examining the dual form of the problem, but were unsuccessful. If the optimal rejection function $r^* \neq r^\delta$ then one can prove several interesting properties, some of which we have proven in Lemma 18, which may be of use in determining bounds on the optimal type II error. However, as the following example illustrates, even determining whether or not the optimal solution outperforms r^δ is not trivial.

Example 5. Let $P = \{0.05, 0.05, \cdots, 0.05, 0.2\}$, $\delta = 0.2$, $\Lambda = 3$ and $D_P(\cdot) = D(\cdot||P)$ be the KL-divergence. Then, solving the linear program gives r^δ (it is possible that other solutions exist, however). Interestingly, changing δ does not appear to change the result (even when taking values as small as $\delta = 0.001$, or as large as $\delta = 0.999$). Furthermore, if we increase Λ to 3.2, we achieve solutions to the linear program which aren't r^δ, but do not improve on its rejection rate (again for the same range of δ values).



(a) Arbitrary (b) Gaussians

Figure 1: Type II error vs. Λ, for $N = 50$ and $\delta = 0.05$. 50 distributions were generated for each value of Λ ($\Lambda = 0.5, 0.1, \cdots, 12.5$). Error bars depict standard error of the mean (SEM).

6.2.1 Numerical Examples

We numerically compare the performance of hard and soft rejection strategies for a constrained game, where $D(Q||P) \geq \Lambda$, for various values of Λ, and two different families of target distributions, P, over a support of size $N = 50$. The families are arbitrary probability mass functions over N events and discretized Gaussians (over N bins). For each Λ we

generated 50 random distributions P for each of the families. For each such P we solved the optimal hard and soft strategies and computed the corresponding worst-case optimal type II error, $1 - \rho(r, Q)$.

Since $\max_Q D(Q||P) = \log(1/\min_i p_i)$, it is necessary that $\min_i p_i \leq 2^{-\Lambda}$ when generating P (to ensure that a Λ-distant Q exists). Distributions in the first family of arbitrarily random distributions, Figure 6.1(a),are generated by sampling a point (p_1) uniformly in $(0, 2^{-\Lambda}]$. The other $N-1$ points are drawn i.i.d. $\sim U(0,1]$, and then normalized so that their sum is $1 - p_1$. The second family, Figure 6.1(b),are Gaussians centered at 0 and discretized over N evenly spaced bins in the range $[-10, 10]$. A (discretized) random Gaussian $N(0, \sigma)$ is selected by choosing σ uniformly in some range $[\sigma_{min}, \sigma_{max}]$. σ_{min} is set to the minimum σ ensuring that the first/last bin will not have "zero" probability (due to limited precision). σ_{max} was set so that the cumulative probability in the first/last bin will be $2^{-\Lambda}$, if possible (otherwise σ_{max} is arbitrarily set to $10 * \sigma_{min}$).

The results for $\delta = 0.05$ are shown in Figure 6.1.Other results (not presented) for a wide variety of the problem parameters (e.g., N, δ) are qualitatively the same. It is evident that both the soft and hard strategies are ineffective for small Λ. Clearly, the soft method has significantly lower error than that of the hard (until Λ becomes "sufficiently large").

7. Low Density Rejection in a Continuous Setting

In Section 5 we presented a number of results on LDRS optimality in a simplified finite and discrete setting. In this section, we reconsider LDRS (now only in the hard setting) in a much more general framework where the learner and adversary distributions are infinitely continuous. After defining this general setting we extend theorem 10 of Section 5 on hard LDRS optimality. The resulting Theorem 30 is obtained by assuming that the adversary strategy space is sufficiently large, now satisfying a continuous extension of Property A called Property $\mathrm{A_{cont}}$ (Property C is not required in the continuous setting).

The main contribution of this section is a reduction of the SCC problem to two-class classification problem. The two-class classification is facilitated by sampling points from a synthetically generated "other class." This other class is generated so that it is uniform over its support, which is appropriately selected around the observed support of P. Using this synthetic sample we obtain a binary training set on which we can train a soft binary classifier. The final δ-valid SCC classifier is then identified by selecting a threshold on the classifier output so as to maximize the type I error up to δ. The entire routine is simple, practical and if the underlying two-class soft classifier learning algorithm runs in $C(n)$ time complexity, our SCC algorithm runs in time $O(C(n) + n)$. An alternative approach where a hard two-class classifier can be used is described by Nisenson (2010).

We show that the SCC routine obtained using this approach is consistent in the sense that if the underlying classification device is consistent then the resulting one-class classifier is asymptotically an LDRF, thus providing an optimal SCC solution when the adversary strategy space satisfies Property $\mathrm{A}_{\mathrm{cont}}$.

7.1 Definitions

The SCC problem in the continuous setting is essentially the same as in the finite case (see Section 2) but now both the source distribution P and the adversary distribution can be infinitely continuous distributions over $\mathbb{R}^d$. Let λ be the Lebesgue measure on $\mathbb{R}^d$. We assume that P is absolutely continuous with respect to λ (in other words, if a Borel set b has zero volume in $\mathbb{R}^d$, then $P(b) = 0$). Denote by p the density function of P and let $\mathrm{supp}(p)$ be its support in $\mathbb{R}^d$.

We define the function $\mathbb{I}_b(x) \triangleq \mathbb{I}(x \in b)$, where $\mathbb{I}(\cdot)$ is the indicator function. For a Borel set b, we define $l_p(b) \triangleq b \bigcup \{x : p(x) = 0\}$.

Definition 26 (Minimum Volume Set). A set $b \subseteq \mathrm{supp}(p)$ is called a minimum volume set of measure $1-\delta$ if $P(b) = 1-\delta$ and for all b' such that $P(b') = P(b) = 1-\delta$, $\lambda(b) \leq \lambda(b')$.

Definition 27 (Low Density Set).

(i) Let $b \subseteq \mathrm{supp}(p)$ be a minimum volume set of measure $1-\delta$. Let m be any set such that $P(m) = \delta$ and $b \bigcap m = \emptyset$. Then, we call m a *core low density set w.r.t. P and δ*,

(ii) Denote by $\mathrm{core}_\delta(P)$ the set of all core low-density sets w.r.t. P and δ.

(iii) We call a set s a *low density set w.r.t. P and δ* if there exists an $m \in \mathrm{core}_\delta(P)$ such that $s = l_p(m)$.

7.2 LDRS optimality in the continuous setting

Definition 28 (Low-Density Rejection Strategy (LDRS) and Function (LDRF)). We define

$$LDRS_\delta(P) \triangleq \left\{ r(\cdot) : \ \exists m \in \mathrm{core}_\delta(P) \ s.t. \ r(\cdot) \equiv \mathbb{I}_{l_P(m)}(\cdot) \right\}.$$

Any function $r(\cdot) \in LDRS_\delta(P)$ is called a δ-tight Low-Density Rejection Function (LDRF), and the Low-Density Rejection Strategy is to choose any δ-tight LDRF.

Definition 29 (Property $\mathbf{A_{cont}}$). We say that two Borel sets j, k satisfy condition $(*)$ if: (i) $j, k \subset \mathrm{supp}(p)$; (ii) $j \cap k = \emptyset$; (iii) $P(j) = P(k)$; and (iv) $\lambda(j) \geq \lambda(k)$.

An adversary strategy space $\mathcal{Q}$ has Property $\mathrm{A}_{\mathrm{cont}}$ w.r.t P, if for every pair j, k satisfying $(*)$: $\forall Q \in \mathcal{Q}$ such that $Q(j) < Q(k)$, $\exists Q' \in \mathcal{Q}$, for which

1. $Q'(j) + Q(j) \geq Q'(k) + Q(k)$;
2. For all Borel sets b for which $b \bigcap (j \bigcup k) = \emptyset$, $Q'(b) = Q(b)$.

The proof of the following theorem can be found in the appendix.

Theorem 30. When the learner is restricted to hard-decisions and $\mathcal{Q}$ satisfies Property $\mathrm{A}_{\mathrm{cont}}$ w.r.t. P, then LDRS is optimal.

7.3 SCC via Two-Class Classification

We propose an SCC routine that relies on a *soft* binary classifier induction. We can use any two-class algorithm, which is consistent in the sense that it minimizes a loss function $\phi(\cdot)$ that is non-negative, differentiable, convex, strictly convex over $[-\infty, 0)$ and satisfies $\phi'(0) < 0$. These conditions are similar but stronger than the conditions required by Bartlett, Jordan, and Mcauliffe (2006), which provide necessary and sufficient conditions for a convex ϕ to be *classification-calibrated*.[7] We note however that the commonly used loss functions as discussed in Bartlett et al. (2006) satisfy our conditions, including the quadratic, truncated-quadratic, exponential and logistic loss functions, to name a few. In the extensions to this section (see Nisenson, 2010) an SCC routine is presented that can utilize any *hard* binary classifier induction algorithm that minimizes either the 0/1, L_1, or hinge loss functions, as well as any of the loss functions defined by Bartlett et al. (2006).[8]

Our SCC algorithm is given a training sample $S_n = \{x_1, \ldots, x_n\}$ of n training examples drawn i.i.d. from an unknown source distribution P over $\mathbb{R}^d$. Given a type-I threshold δ the algorithm outputs a hard rejection function $r(\cdot)$ over $\mathbb{R}^d$. The main idea of the algorithm is based on the following observation. If our domain is bounded, we can define a two-class classification problem where the first class is P and the other class is a uniform distribution over the (bounded) domain. Then, the output of a consistent soft binary classifier is strictly monotonically increasing with $p(\cdot)$ (the density of P) over the support of P (it is only weakly monotone in $p(\cdot)$ over the whole domain). Therefore, thresholding the classifier's output, with an appropriate quantile, identifies a δ-valid level-set in P, inducing a rejection function.

In practice, sampling from a uniform distribution over large domains is computationally hard and moreover, undefined for unbounded domains. Our algorithm avoids these obstacles by sampling uniformly in grid cells containing sampled points from P. An additional complication arises in cases where the density p is flat over some regions, which results in discontinuities of the level sets. This is a known issue in level set estimation and

7. Our additional conditions are differentiability everywhere and strict convexity over $[-\infty, 0)$. The reason for these extra conditions is that we threshold the soft classifier's output and don't merely use its sign for classification.
8. The use of a hard classifier (as opposed to a soft one) results in a time complexity penalty of a factor of $O(\log n)$.

is often avoided by assuming that there are no flat regions in p, in particular in regions corresponding to the δ level set (Tsybakov, 1997; Molchanov, 1990). We don't assume this; our algorithm handles flat regions in p by jittering the classifier output using a small and vanishing (in n) random noise (see step 6 in the algorithm below). The resulting algorithm is computationally efficient and practical.

A major component of our algorithm is determining a threshold by quantile estimation. This occurs in Step 7 of the algorithm. We apply a known estimator (Uhlmann, 1963; Zieliński, 2004) that is unbiased and has certain optimal characteristics (see below). This quantile estimator assumes that the cumulative distribution function (cdf), F, underlying the sample, is continuous, and is defined over $\mathbb{R}$ (i.e., F is the cdf of a real random variable). Let t_μ be the estimate of the μ-quantile of F, given n sample points drawn i.i.d. according to F. The estimator is *unbiased* if $\mathbf{E}_F[F(t_\mu)] = \mu$. Its variance is $Var_F[F(t_\mu)]$. The estimator we use is called the "uniformly minimum variance unbiased estimator." It was introduced by Uhlmann (1963) and we rely on analysis by Zieliński (2004). This estimator can only be used for estimating μ-quantiles that satisfy $\frac{1}{n+1} \leq \mu \leq \frac{n}{n+1}$, which is equivalent to requiring that $n \geq \max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}$. The estimator chooses an index π_μ in $[1, \ldots, n]$, and the estimate of the μ-quantile is the π_μ-th order statistic; in other words, if our sample points are sorted in increasing order, then the estimate is the π_μ-th element. π_μ is calculated as follows:

- Set $k \stackrel{\triangle}{=} \lfloor (n+1)\mu \rfloor$.
- Set $\beta \stackrel{\triangle}{=} (n+1)\mu - k$.
- With probability β, set $\pi_\mu = k+1$, and with probability $1-\beta$, set $\pi_\mu = k$.

The estimator's variance is (Zieliński, 2004):

$$\frac{\beta(1-\beta)}{(n+1)(n+2)} + \frac{\mu(1-\mu)}{n+2}.$$

The variance is maximized when $\beta = \mu = \frac{1}{2}$, and thus the variance is at most $\frac{1}{4(n+1)}$. Moreover, according to Zieliński (2004), the estimator is unbiased and its variance is not greater than that of any other unbiased estimator within the family of estimators that can be defined using a probability distribution over single order statistics. For very small samples with $n < \max\left\{\frac{\mu}{1-\mu}, \frac{1-\mu}{\mu}\right\}$, we "fall-back" to a simple "default" estimator, which sets $\pi_\mu \stackrel{\triangle}{=} \lceil n\mu \rceil$. We term this quantile-estimation algorithm the "uniformly minimum variance unbiased (with fall-back) estimator," or the "UMVUFB estimator."

The algorithm is as follows:

1. Define a grid over $\mathbb{R}^d$ with arbitrary origin and positive cell side length g_n. Let $g_n \to 0$, be such that $ng^d \to \infty$. For example, $g_n \stackrel{\triangle}{=} n^{-\frac{1}{d+2}}$. Select an arbitrary

origin x_0, for example, uniformly at random from the unit-hypercube. For any point $x = (x^{(1)}, \ldots, x^{(d)})$, define the function

$$A_n(x) \triangleq \left\lfloor \frac{x - x_0}{g_n} \right\rfloor \triangleq \left(\left\lfloor \frac{x^{(1)} - x_0^{(1)}}{g_n} \right\rfloor, \left\lfloor \frac{x^{(2)} - x_0^{(2)}}{g_n} \right\rfloor, \ldots, \left\lfloor \frac{x^{(d)} - x_0^{(d)}}{g_n} \right\rfloor \right).$$

For each point x, $A_n(x)$ specifies the coordinates of the "lower left" corner of the grid cell containing x.

2. Define the set $G_P^n = \bigcup_{x \in S_n} A_n(x)$ of covered grid cell corners.

3. Generate an artificial sample O_n of size n from the "other class." Each point is selected independently at random as follows:

   (a) Choose $a \in G_P^n$ uniformly at random.

   (b) Choose a point x uniformly at random from the unit-hypercube.

   (c) The new artificial sample point is $o \triangleq a + g_n \cdot x$.

4. Using the training sample consisting of S_n (labeled $+1$) and O_n (labeled -1), train a soft binary classifier $h_n(\cdot)$.

5. Define a confidence margin for the δ threshold. Select any $\theta_n \to \infty$ such that $\theta_n = o(\sqrt{n})$, for example, take $\theta_n \triangleq \sqrt[3]{n}$. Now define $\delta_n^+ \triangleq \delta + \frac{1}{\theta_n}$. Choose $\delta_n^- \leq \delta - \frac{1}{\theta_n}$ be such that $\delta_n^- \to \delta$.

6. Jitter the classifier output. Let X_P be a random variable where $X_P \sim P$ and $Y_n \triangleq h_n(X_p)$. Let $\Phi(\cdot)$ be the cumulative distribution function of $N(0,1)$, and let m_n be such that $\frac{\Phi(-m_n)}{\Phi(m_n)} = o\left(\frac{1}{\theta_n}\right)$, for example $m_n \triangleq e^n$. Let $\sigma_n \triangleq o\left(\frac{1}{m_n}\right)$, for example, $\sigma_n \triangleq \frac{e^{-n}}{m_n} = e^{-2n}$. Let $\varepsilon \sim N[0, \sigma_n^2]$, and set $Z_n \triangleq Y_n + \varepsilon$.

7. We use the following threshold mechanism. We will select two thresholds t_n^- and t_n^+ on Z_n. The cutoff is always t_n^- and it is inclusive when $t_n^- < t_n^+$. Specifically, let t_n^- and t_n^+ be estimates of the $\left(\Phi(m_n)\delta_n^- + \frac{\Phi(-m_n)}{2}\right)$-quantile and $\left(\Phi(m_n)\delta_n^+ + \frac{\Phi(-m_n)}{2}\right)$-quantile of Z_n, respectively. In order to establish these estimates we require a sample from Z_n. The following procedure produces a list of sample points S_Z.

   - Set $S_Z = []$, i.e. S_Z is an empty list.
   - For each $x \in S_n$: Choose a value $\epsilon_x \sim N[0, \sigma_n^2]$ and append the value $h_n(x) + \epsilon_x$ onto S_Z.

   The sample S_z is then the input to the UMVUFB estimator defined above.

8. Define the rejection function

$$r_n(x) \triangleq \begin{cases} 1 & A_n(x) \notin G_P^n; \\ \mathbb{I}(h_n(x) \leq t_n^-) & A_n(x) \in G_P^n \text{ and } t_n^- < t_n^+; \\ \mathbb{I}(h_n(x) < t_n^-) & \text{otherwise.} \end{cases}$$

Remark 31. Instead of a soft classifier, $h_n(\cdot)$ could have been any consistent class-probability estimator, where $h_n(x)$ is the estimate of $\Pr\{+1|x\}$. See (Nisenson, 2010) for details. $h_n(\cdot)$ could also be a consistent ranking algorithm (see, e.g., Clémençon, Lugosi, & Vayatis, 2005). In this case, the quantile estimator must select a single sample point to represent the quantile. All comparison operations (e.g. $<, \leq$), including those done by the quantile estimator, must be performed by the ranking algorithm. The ranking algorithm must also be able to distinguish between $t_n^- < t_n^+$ and $t_n^- = t_n^+$.

Let U_n (with density u_n) be the distribution of O_n (defined in Step 3). Clearly, u_n is uniform over its bounded support. As previously noted, if the support of the generated distribution is significantly larger than that of P, an exorbitant number of points may need to be generated in practice in order to reject low density areas in P (Davenport et al., 2006). The following lemma shows that the probability of generating points outside of p's support, almost surely tends to zero.

Lemma 32. $U_n(\mathbb{R}^d \setminus \text{supp}(p)) \xrightarrow{\text{a.s.}} 0$.

Proof Recall that g_n is a sequence of positive numbers such that $\lim_{n\to\infty} ng_n^d = \infty$ and $\lim_{n\to\infty} g_n = 0$. Define a sequence of positive numbers g'_n, such that $g'_n \geq 2g_n$, $g'_n \to 0$ and $\lim_{n\to\infty} \frac{ng'^d_n}{\log n} = \infty$. Define $A(x, g'_n) \triangleq \{y : y \in \mathbb{R}^d \text{ and } ||x - y||_\infty \leq g'_n\}$. Define $T_n \triangleq \bigcup_{i=1}^n A(x_i, g'_n)$. Devroye and Wise (1980) show that for any probability measure ν on the Borel sets of $\mathbb{R}^d$ whose restriction to $\text{supp}(p)$ is absolutely continuous w.r.t. P, it holds that $\nu(T_n \,\Delta\, \text{supp}(p)) \xrightarrow{\text{a.s.}} 0$. We note that the grid cell of x is always a sub-region of $A(x, g'_n)$, and therefore $\text{supp}(u_n) \subseteq T_n$. Thus, noting that $U_n(\mathbb{R}^d \setminus \text{supp}(p)) = U_n(\text{supp}(u_n) \setminus \text{supp}(p))$,

$$\begin{aligned} U_n(\text{supp}(u_n) \setminus \text{supp}(p)) &\leq \sup_m U_m(\text{supp}(u_n) \setminus \text{supp}(p)) \\ &\leq \sup_m U_m(T_n \setminus \text{supp}(p)) \\ &\leq \sup_m U_m(T_n \,\Delta\, \text{supp}(p)) \\ \Rightarrow \lim_{n\to\infty} U_n(\text{supp}(u_n) \setminus \text{supp}(p)) &\leq \lim_{n\to\infty} \sup_m U_m(T_n \,\Delta\, \text{supp}(p)) \\ &= \sup_m \lim_{n\to\infty} U_m(T_n \,\Delta\, \text{supp}(p)) \\ &\overset{\text{a.s.}}{=} 0. \end{aligned}$$

■

Remark 33. It is difficult to establish exact convergence rates in Lemma 32 without constraints on P. For cases where $\lambda(\mathbb{R}^d \setminus \text{supp}(p)) = 0$, we obviously have that $U_n(\mathbb{R}^d \setminus \text{supp}(p)) = 0$. This is the case, for example, for finite mixtures of Gaussians.

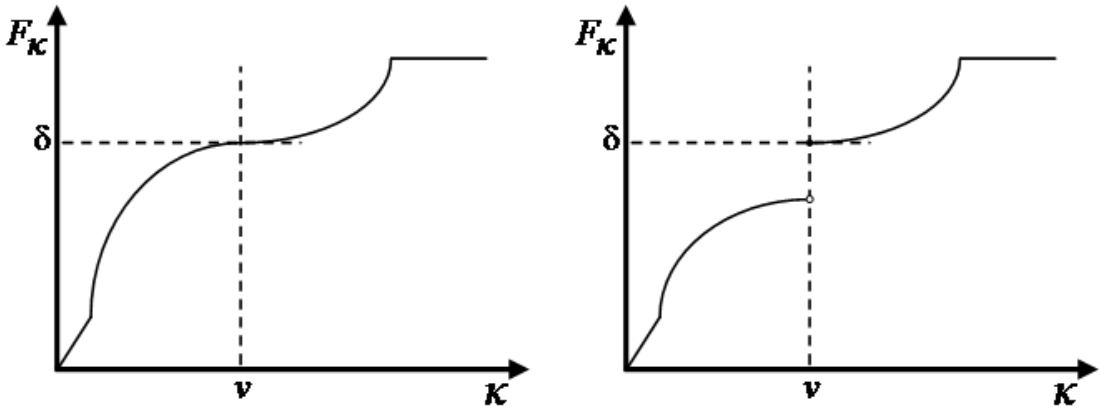
If there exists a constant K, such that $|p(x) - p(y)| \leq K$ whenever $||x - y||_\infty \leq g_n$, we can establish an upper bound on the rate. The condition $||x - y||_\infty \leq g_n$ is equivalent to x and y being in the same grid-cell. Therefore, if $p(x) > K$, then for all y in the same grid-cell, $p(y) > 0$. Note that if p is Lipschitz continuous such that $|p(x) - p(y)| \leq \frac{K}{g_n}||x - y||_\infty$, then p meets the above condition. Let $1 - \eta$ be a desired confidence level. Let C_K^η be the number of cells in the grid which contain more than $Kng_n^d + \sqrt{\frac{n(\ln|G_p^n| - \ln\eta)}{2}}$ sample points. Then, using Hoeffding's inequality, it isn't hard to show that with probability at least $1-\eta$, $U_n(\mathbb{R}^d \setminus \text{supp}(p)) \leq 1 - \frac{C_K^\eta}{|G_P^n|}$.

Definition 34 (Quantile). Let ξ be a random variable whose domain is in $\mathbb{R}$. We say that t is a *μ-quantile* of ξ if

$$t \in S_\xi(\mu) \triangleq \{\tau \in \mathbb{R} : \Pr\{\xi < \tau\} \leq \mu \text{ and } \Pr\{\xi \leq \tau\} \geq \mu\}.$$

Define a new random variable κ to represent the level sets of P. Formally, its cumulative distribution function is $F_\kappa(t) = P(\{x : p(x) \leq t\})$.

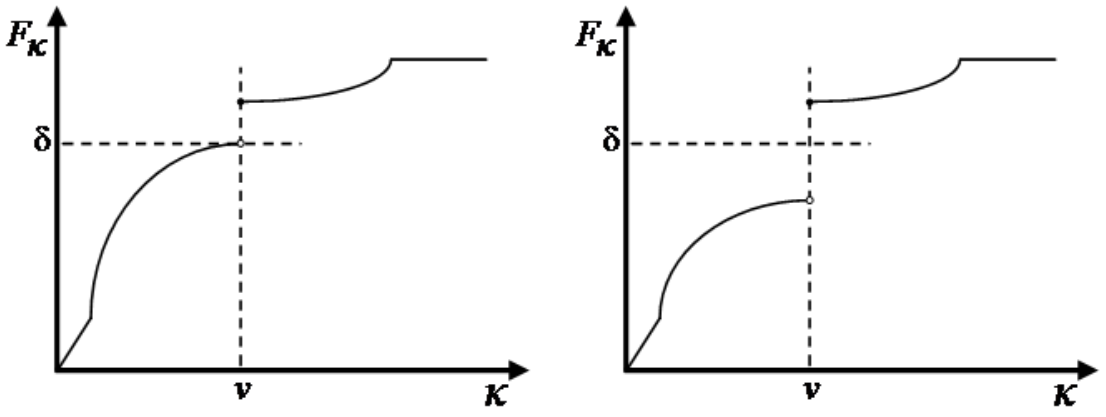
Definition 35. Let v be any δ-quantile of κ. We say P has a *δ-jump* if $F_\kappa(v) > \delta$.



(a) (b)

Figure 2: The cumulative distribution function F_κ when P doesn't have a δ-jump.

We now will consider two cases, one where P doesn't have a δ-jump and one where it does. See Figure 2 and Figure 3. In all figures the (unique) δ-quantile of κ is marked by



(a) (b)

Figure 3: The cumulative distribution function F_κ when P does have a δ-jump.

v. Note that a quantile need not be unique, in particular there will be a range of values wherever F_κ is flat.

Definition 36. A rejection function $r(\cdot)$ is called a *δ-maximal level-set estimator for P* if, for some $v \in S_\kappa(\delta)$, either:

1. P doesn't have a δ-jump and $r(x) \equiv \mathbb{I}(p(x) \leq v)$, almost everywhere.
2. P has a δ-jump and $r(x) \equiv \mathbb{I}(p(x) < v)$, almost everywhere.

Note that if P doesn't have a δ-jump, then a δ-maximal level-set estimator for P is a δ-tight LDRF. We will now prove that the output of the algorithm is asymptotically (almost surely) a δ-maximal level-set estimator for P.

Theorem 37. Let $\{U'_n\}$, $n = 1, 2, \ldots$, be a sequence of probability measures such that for each n, U'_n has uniform density u'_n over its bounded support, and $\lim_{n\to\infty} P(\mathrm{supp}(u'_n)) = 1$. Define a Bayesian binary classification problem for each n. Let the first class, $c_1 \equiv +1$ have distribution P, and the second class $c_2 \equiv -1$ have distribution U'_n. The classes' prior probabilities are $\Pr\{+1\} = \Pr\{-1\} = \frac{1}{2}$. Let $\phi(\cdot)$ be a non-negative, differentiable, convex loss function such that it is strictly convex on $[-\infty, 0)$ and $\phi'(0) < 0$. Let $h^*_n(\cdot)$ be the soft Bayes-optimal classifier that minimizes the expected loss. Define a random variable $Y^*_n \triangleq h^*_n(X_P)$. Let t^*_n be a δ-quantile of Y^*_n. Define the rejection function:

$$r^*_n(x) \triangleq \begin{cases} 1 & x \notin \mathrm{supp}(u'_n); \\ \mathbb{I}(h^*_n(x) \leq t^*_n) & x \in \mathrm{supp}(u'_n) \text{ and } P \text{ doesn't have a } \delta\text{-jump}; \\ \mathbb{I}(h^*_n(x) < t^*_n) & \text{otherwise.} \end{cases}$$

Then, $r^*(\cdot) \triangleq \lim_{n\to\infty} r^*_n(\cdot)$ is a δ-maximal level-set estimator for P.

Proof We first consider $x \in \text{supp}(u'_n)$. Define the function $\psi_n(x) \triangleq \frac{p(x)}{u'_n(x)}$, defined over $\text{supp}(u'_n)$. From Bayes theorem, it is not hard to show that $\Pr\{+1|x\} = \frac{p(x)}{p(x)+u'_n(x)}$. The loss for a point x when we assign it value y is (Bartlett et al., 2006):

$$\begin{aligned}\ell(x,y) &\triangleq \Pr\{+1|x\}\phi(y) + \Pr\{-1|x\}\phi(-y) \\ &= \frac{p(x)\phi(y) + u'_n(x)\phi(-y)}{p(x) + u'_n(x)}.\end{aligned}$$

It is easy to verify that for a fixed x, at the minimum (over y) of $\ell(x,y)$, $p(x)\phi'(y) = u'_n(x)\phi'(-y)$. Alternatively: $\phi'(-y) = \psi_n(x)\phi'(y)$. Let x_1 and x_2 be two points such that $\psi_n(x_1) > \psi_n(x_2)$. Note that $\min\{\phi'(y), \phi'(-y)\} \leq \phi'(0) < 0$ for all y. Let $c_i \triangleq \psi_n(x_i)$ and y_i be a solution to $\phi'(-y) = c_i\phi'(y)$, for $i \in \{1,2\}$. Note that $c_1, c_2 \geq 0$ and therefore, in order for equality to occur it is necessary that $\phi'(y_i), \phi'(-y_i) \leq 0$ (with equality only if $c_i = 0$). We can now rewrite $\phi'(-y_i) = c_i\phi'(y_i)$ as $|\phi'(-y_i)| = c_i|\phi'(y_i)|$.

We will now prove that $y_1 > y_2$. Assume by contradiction that the statement is false. Then $y_2 \geq y_1$. Therefore, $|\phi'(y_2)| \leq |\phi'(y_1)|$ and $|\phi'(-y_2)| \geq |\phi'(-y_1)|$. Since $\psi_n(x_1) > \psi_n(x_2)$, it follows that $c_1 > c_2$. If $c_2 = 0$, then $0 = |\phi'(-y_2)| \geq |\phi'(-y_1)|$. Therefore $\phi'(-y_1) = 0$ and $\phi'(y_1) < 0$, which gives $0 = |\phi'(-y_1)| = c_1|\phi'(y_1)| < 0$, which is a contradiction. Thus, $c_2 \neq 0$, and $|\phi'(-y_2)| = c_2|\phi'(y_2)| \leq c_2|\phi'(y_1)| = \frac{c_2}{c_1}|\phi'(-y_1)| \leq \frac{c_2}{c_1}|\phi'(-y_2)| < |\phi'(-y_2)|$. Contradiction.

Now consider the case where $c_i = \frac{|\phi'(-y_i)|}{|\phi'(y_i)|} > 0$. Therefore, $\phi'(y_i), \phi'(-y_i) < 0$. Since $\phi(\cdot)$ is strictly convex over $[-\infty, 0)$ it follows that as y_i increases $|\phi'(y_i)|$ decreases and $|\phi'(-y_i)|$ increases. Therefore, if $\psi_n(x_1) = \psi_n(x_2) > 0$, there is a unique solution.

Therefore, $h^*_n(x)$ is monotonically increasing with $\psi_n(x)$, almost everywhere over $\text{supp}(u_n)$ and strictly monotonically increasing with $\psi_n(x)$, almost everywhere over $\text{supp}(u_n) \bigcap \text{supp}(p)$. Since $u'_n(\cdot)$ is constant over its support, this implies: $p(x_1) < p(x_2) \Rightarrow h^*_n(x_1) < h^*_n(x_2)$, and $0 < p(x_1) = p(x_2) \Rightarrow h^*_n(x_1) = h^*_n(x_2)$. Therefore, for some v^*_n, $\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : h^*_n(x) \leq t^*_n\}$ is identical to $\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : p(x) \leq v^*_n\}$ (with the possible exception of a set of points of zero Lebesgue measure). Recalling that $Y^*_n = h^*_n(X_P)$ for $X_P \sim P$

and that $P(\text{supp}(u'_n)) \to 1$:

$$\begin{aligned}
\lim_{n\to\infty} t^*_n &\in \lim_{n\to\infty} \{\tau \in \mathbb{R} : \Pr\{Y^*_n < \tau\} \le \delta \text{ and } \Pr\{Y^*_n \le \tau\} \ge \delta\} \\
&= \lim_{n\to\infty} \{\tau \in \mathbb{R} : P(\{x : h^*_n(x) < \tau\}) \le \delta \text{ and } P(\{x : h^*_n(x) \le \tau\}) \ge \delta\} \\
&= \lim_{n\to\infty} \{\tau \in \mathbb{R} : P(\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : h^*_n(x) < \tau\}) \le \delta \text{ and} \\
&\qquad\qquad P(\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : h^*_n(x) \le \tau\}) \ge \delta\} \\
\Rightarrow \lim_{n\to\infty} v^*_n &\in \lim_{n\to\infty} \{\tau' \in \mathbb{R} : P(\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : p(x) < \tau'\}) \le \delta \text{ and} \\
&\qquad\qquad P(\{x \in \text{supp}(u'_n) \bigcap \text{supp}(p) : p(x) \le \tau'\}) \ge \delta\} \\
&= \lim_{n\to\infty} \{\tau' \in \mathbb{R} : P(\{x : p(x) < \tau\}) \le \delta \text{ and } P(\{x : p(x) \le \tau'\}) \ge \delta\} \\
&= \{\tau' \in \mathbb{R} : P(\{x : p(x) < \tau\}) \le \delta \text{ and } P(\{x : p(x) \le \tau'\}) \ge \delta\} \\
&= \{\tau' \in \mathbb{R} : \Pr\{\kappa < \tau'\} \le \delta \text{ and } \Pr\{\kappa \le \tau'\} \ge \delta\} \\
&= S_\kappa(\delta)
\end{aligned}$$

Therefore, let $v^\delta_p \in S_\kappa(\delta)$ be such that $v^\delta_p = \lim_{n\to\infty} v^*_n$. Note that since $\delta > 0$, $v^\delta_p > 0$ (otherwise $\delta \le P(\{x : p(x) \le v^\delta_p) = 0)$. Therefore, for sufficiently large n, $v^*_n > 0$.

Let us assume that P doesn't have a δ-jump. Therefore, for almost every $x \in \text{supp}(u'_n)$, $\mathbb{I}(h^*_n(x) \le t^*_n) = \mathbb{I}(p(x) \le v^*_n)$. Then almost everywhere in $\text{supp}(u'_n)$: $r(x) = \lim_{n\to\infty} \mathbb{I}(h^*_n(x) \le t^*_n) = \mathbb{I}(p(x) \le v^\delta_p)$. It is given that $P(\mathbb{R}^d \setminus \text{supp}(u'_n)) \to 0$. Therefore, $\lambda(\{x \notin \text{supp}(u'_n) : p(x) > v^\delta_p\}) \to 0$. which is equivalent to $\lambda(\{x \notin \text{supp}(u'_n) : \mathbb{I}(p(x) \le v^\delta_p) \ne r^*_n(x)\}) \to 0$.

If P has a δ-jump, the proof is almost identical, only with minor changes in the strengths of inequalities. For almost every $x \in \text{supp}(u'_n)$: $r(x) = \lim_{n\to\infty} \mathbb{I}(h^*_n(x) < t^*_n) = \mathbb{I}(p(x) < v^\delta_p)$, and $\lambda(\{x \notin \text{supp}(u'_n) : \mathbb{I}(p(x) < v^\delta_p) \ne r^*_n(x)\}) \to 0$. ■

We will now make clear the relation between the algorithm given and Theorem 37. Clearly $\{U_n\}$ is a series of distributions each having a uniform density, u_n, over its bounded support. We will now prove that $P(\text{supp}(u_n)) \xrightarrow{\text{a.s.}} 1$.

Lemma 38. For any $\epsilon > 0$, $\Pr\{P(\text{supp}(u_n)) \le 1-\epsilon\} \le 2e^{-2n\epsilon^2 - no(1)}$ and $P(\text{supp}(u_n)) \xrightarrow{\text{a.s.}} 1$.

Proof We define $G(x)$ to be the cell in the grid which contains x. Define $c(b) \triangleq |\{S \bigcap b\}|$ to be the count of the number of training samples which fall within set b. Then the histogram density estimate is $\tilde{p}_n(x) \triangleq \frac{c(G(x))}{nh^d}$. As shown by Devroye and Gyorfi (2002) in Theorem 5.6, $\Pr\left\{\int_{\mathbb{R}^d} |p(x) - \tilde{p}_n(x)|\lambda(dx) > 2\epsilon\right\} \le 2e^{-2n\epsilon^2 - no(1)}$. However, since P is absolutely continuous w.r.t. λ, it follows from Scheffé's theorem (Scheffé (1947), used as Theorem 5.4 by Devroye and Gyorfi (2002)), that for any Borel set B over $\mathbb{R}^d$, $\Pr\left\{\int_B |p(x) - \tilde{p}_n(x)|\lambda(dx) > \epsilon\right\} \le 2e^{-2n\epsilon^2 - no(1)}$.

By definition, $\tilde{p}_n(x) = 0$ for all $x \notin \text{supp}(u_n)$. Therefore:

$$\begin{aligned}\Pr\left\{P\left(\mathbb{R}^d \setminus \text{supp}(u_n)\right) > \epsilon\right\} =& \Pr\left\{\int_{\mathbb{R}^d \setminus \text{supp}(u_n)} |p(x) - \tilde{p}_n(x)|\lambda(dx) > \epsilon\right\} \\ \leq& 2e^{-2n\epsilon^2 - no(1)}.\end{aligned}$$

Since this is true for any ϵ, it immediately follows that $\Pr\{\lim_{n\to\infty} P(\mathbb{R}^d \setminus \text{supp}(u_n)) \neq 0\} = 0$, or $P(\text{supp}(u_n)) \xrightarrow{\text{a.s.}} 1$. ■

Therefore, the only remaining part is to show how t_n^- and t_n^+ relate to t_n^* and to whether P has a δ-jump or not. We note that for all x, at the limit, $h_n^*(x) = h_n(x)$ and therefore, $Y_n^* \equiv Y_n$ (i.e. they are distributed identically). For sufficiently large n, the quantile estimator used is unbiased with standard deviation vanishing at a rate of $O\left(\frac{1}{\sqrt{n+1}}\right)$ (Zieliński, 2004). Therefore, since $\theta_n = o(\sqrt{n})$, it follows that $\frac{1}{\sqrt{n}} = o(\frac{1}{\theta_n})$, and thus for sufficiently large n, t_n^- and t_n^+ are tightly concentrated around a $\left(\Phi(m_n)\delta_n^- + \frac{\Phi(-m_n)}{2}\right)$-quantile (which is not greater than the $\left(\Phi(m_n)\left[\delta - \frac{1}{\theta_n}\right] + \frac{\Phi(-m_n)}{2}\right)$-quantile) and a $\left(\Phi(m_n)\left[\delta + \frac{1}{\theta_n}\right] + \frac{\Phi(-m_n)}{2}\right)$-quantile for Z_n, respectively. Therefore, since $\frac{\Phi(-m_n)}{\Phi(m_n)} = o\left(\frac{1}{\theta_n}\right)$, by the following lemma, t_n^- and t_n^+ are also tightly concentrated around a δ_n^--quantile and a δ_n^+-quantile for Y_n, respectively.

Lemma 39. Let $m \geq 0$. Let t_μ be a $\left(\Phi(m)\mu + \frac{\Phi(-m)}{2}\right)$-quantile of Z_n. Then, for some μ' such that $|\mu' - \mu| \leq \frac{\Phi(-m)}{2\Phi(m)}$, a μ'-quantile of Y_n, $t_{\mu'}^*$, satisfies $|t_{\mu'}^* - t_\mu| \leq m\sigma_n$.

Proof

$$\begin{aligned}\mu + \frac{\Phi(-m)}{2\Phi(m)} \geq& \frac{1}{\Phi(m)}\Pr\{Z_n < t\} = \frac{1}{\Phi(m)}\Pr\{Y_n < t + \varepsilon\} \\ \geq& \frac{1}{\Phi(m)}\Pr\{Y_n < t - m\sigma_n\}\Pr\{\varepsilon \geq -m\sigma_n\} = \Pr\{Y_n < t - m\sigma_n\}\end{aligned}$$

$$\begin{aligned}\mu + \frac{\Phi(-m)}{2\Phi(m)} \leq& \frac{1}{\Phi(m)}\Pr\{Z_n \leq t\} = \frac{1}{\Phi(m)}\Pr\{Y_n \leq t + \varepsilon\} \\ \leq& \frac{1}{\Phi(m)}\left[\Pr\{Y_n \leq t + m\sigma_n\}\Pr\{\varepsilon \leq m\sigma_n\} + \Pr\{\varepsilon > m\sigma_n\}\right] \\ \leq& \frac{1}{\Phi(m)}\left[\Pr\{Y_n \leq t + m\sigma_n\}\Phi(m) + \Phi(-m)\right] \\ =& \Pr\{Y_n \leq t + m\sigma_n\} + \frac{\Phi(-m)}{\Phi(m)}.\end{aligned}$$

Therefore, $\mu + \frac{\Phi(-m)}{2\Phi(m)} \geq \Pr\{Y_n < t - m\sigma_n\}$ and $\mu - \frac{\Phi(-m)}{2\Phi(m)} \leq \Pr\{Y_n \leq t + m\sigma_n\}$. Let $\Delta_1 \triangleq \Pr\{Y_n < t - m\sigma_n\} - \mu$, and let $\Delta_2 \triangleq \Pr\{Y_n \leq t + m\sigma_n\} - \mu$. Note that $t - m\sigma_n$ is a

$(\mu+\Delta_1)$-quantile and that $t+m\sigma_n$ is a $(\mu+\Delta_2)$-quantile of Y_n. Therefore, since $\Delta_2 \geq \Delta_1$, for every $\Delta \in [\Delta_1, \Delta_2]$, there is some t' such that $|t'-t_\mu| \leq m\sigma_n$, which is a $(\mu+\Delta)$-quantile of Y_n. To complete the proof, note that $\Delta_1 \leq \frac{\Phi(-m)}{2\Phi(m)}$ and $\Delta_2 \geq -\frac{\Phi(-m)}{2\Phi(m)}$. Therefore, there exists some $\Delta \in \left[-\frac{\Phi(-m)}{2\Phi(m)}, \frac{\Phi(-m)}{2\Phi(m)}\right]$ such that $\Delta \in [\Delta_1, \Delta_2]$. ■

Theorem 40. The rejection function output by the algorithm is (almost surely) identical to that of Theorem 37 at the limit, where $U'_n \triangleq U_n$.

Proof By definition, $x \in \text{supp}(u'_n) \Leftrightarrow A_n(x) \in G_p^n$.

We represent by t_n^{*-} and t_n^{*+} the δ_n^- and δ_n^+ quantiles of Y_n^*, around which (for sufficiently large n), t_n^- and t_n^+ are tightly concentrated. In particular, $t_n^{*-} < t_n^{*+} \Rightarrow t_n^- < t_n^+$. Note that $t_n^- \leq t_n^+$ and $t_n^{*-} \leq t_n^* \leq t_n^{*+}$ always, and at the limit, $t_n^- = t_n^{*-} = t_n^* = t_n^{*+} = t_n^+$. We now consider four cases.

In the first, P doesn't have a δ_n^--jump or a δ-jump (see Figure 7.1(a)).Then, $t_n^{*-} < t_n^* < t_n^{*+}$. Therefore, for $x \in \text{supp}(u'_n)$, at the limit: $\mathbb{I}(h_n(x) \leq t_n^-) = \mathbb{I}(h_n^*(x) \leq t_n^*(x))$.

In the second, P has a δ_n^--jump but it doesn't have a δ-jump (see Figure 7.1(b)).Then, for sufficiently large n, $t_n^{*-} = t_n^* < t_n^{*+}$ and $\delta_n^- < \Pr\{Y_n^* \leq t_n^{*-}\} \leq \delta$. Therefore, for $x \in \text{supp}(u'_n)$, at the limit: $\mathbb{I}(h_n(x) \leq t_n^-) = \mathbb{I}(h_n^*(x) \leq t_n^*(x))$.

In the third, P doesn't have a δ_n^--jump but it does have a δ-jump (see Figure 7.2(a)).Then, for sufficiently large n, $t_n^{*-} < t_n^* = t_n^{*+}$. Therefore, for $x \in \text{supp}(u'_n)$, at the limit: $\mathbb{I}(h_n(x) \leq t_n^-) = \mathbb{I}(h_n^*(x) < t_n^*(x))$.

In the fourth, P has both a δ_n^--jump and a δ-jump (see Figure 7.2(b)).Then, for sufficiently large n, $t_n^{*-} = t_n^* = t_n^{*+}$. Therefore, for $x \in \text{supp}(u'_n)$, at the limit: $\mathbb{I}(h_n(x) < t_n^-) = \mathbb{I}(h_n^*(x) < t_n^*(x))$. ■

Remark 41 (Rates of Convergence and Finite Sample Notes). The time complexity for our algorithm is $O(C(n)+n)$, where $C(n)$ is the time complexity for the soft-classification algorithm. The rate of convergence for the given algorithm is $\Theta\left(\frac{1}{\theta_n}\right) = \Theta\left(\frac{1}{n^{\frac{1}{2}+\epsilon}}\right)$, for any $\epsilon > 0$, in addition to the classifier's rate of convergence.[9] θ_n is only affected by the quantile-estimator used. In our case, the quantile estimator utilized only requires that F, the cdf whose quantile is being estimated, be continuous. To meet this condition we added the

9. The classifier doesn't truly need to minimize the loss. Depending on the quantile-estimator, it is possible that only classifier errors which result in "ordering violations" across the δ-quantile can affect the output (beyond whether a strong or weak inequality is used for testing the threshold). Thus, faster rates than the classifier's convergence rate to the minimum may be possible. Also, ranking algorithms (see, e.g., Clémençon et al., 2005) could be used instead of soft-classification. In this case, achievable error rates could provide (loose) upper bounds on such ordering violations.

noise term ε. Note that ε has no effect on the convergence rate; this is because σ_n can vanish as fast as desired. Similarly, by Lemma 39, we can achieve arbitrarily tight bounds on the nearness of the quantiles of Z_n and Y_n by increasing the rate at which m_n tends to infinity.

For finite sample sizes, some additional modifications are advisable. First, in order to ensure that $h_n^*(\cdot) = \mathbf{E}_{U_n}[h_n(\cdot)]$, O_n should be of size $n' \sim NB(n, \frac{1}{2})$, and not n. It is also possible to use a non-uniform prior probability (without affecting the algorithm's correctness), if it is desired. A validation set could be used for determining the quantile-estimates, rather than the training set. Note that for finite samples, it is not guaranteed that $\Pr\{Y_n \leq t_n^-\} \approx \delta_n^-$. In fact, it is possible to be significantly larger if Y_n has a large jump in the range $[t_n^- - m\sigma_n, t_n^-)$. Since by Lemma 39, $\Pr\{Y_n \leq t_n^- - m\sigma_n\} \leq \delta_n^- + \frac{\Phi(-m)}{2\Phi(m)} + o\left(\frac{1}{\theta_n}\right)$, almost surely, we can address this issue by refining the definition of the rejection function output by the algorithm:

$$r_n(x) \triangleq \begin{cases} 1 & A_n(x) \notin G_P^n; \\ \mathbb{I}(h_n(x) \leq t_n^- - m\sigma_n) & A_n(x) \in G_P^n \text{ and } t_n^- < t_n^+; \\ \mathbb{I}(h_n(x) < t_n^- - m\sigma_n) & \text{otherwise.} \end{cases}$$

Note that this fix isn't possible when using a ranking algorithm in place of a soft binary classifier, since only points, and not values, can be compared (i.e., x and the chosen quantile point in S_Z are compared in order to determine whether to reject x).

Finally, one needs to determine δ_n^-, so that it is guaranteed (with high probability) that $\rho(r_n, P) \leq \delta$. To accomplish this, one must take into account the quantile estimator used, since $\delta_n^- \leq \delta - \frac{1}{\theta_n}$, and $P(\mathbb{R}^d \setminus \text{supp}(u_n))$, since this is always rejected. It is known (Hall & Hannan, 1988) for the histogram density estimator, upon which the sampling of O_n in the algorithm is loosely-based, that g_n of order $n^{-\frac{1}{d+2}}$ is optimal for minimizing L_b distance for $1 \leq b < \infty$, and that g_n of order $\left(\frac{\log n}{n}\right)^{\frac{1}{d+2}}$ is the correct order for minimizing L_∞ distance. However, we are only interested in $P(\mathbb{R}^d \setminus \text{supp}(u_n))$. We note that this is just the missing mass. Let n_1 be the number of grid cells containing exactly one point from the sample. Then, as shown by Robert and Schapire (2000), with probability at least $1 - \eta$, $P(\mathbb{R}^d \setminus \text{supp}(u_n)) \leq \frac{n_1}{n} + \left(2\sqrt{2} + \sqrt{3}\right)\sqrt{\frac{\ln 3 - \ln \eta}{n}}$. Clearly, increasing g_n results in n_1 decreasing. Therefore, g_n should be large in order to minimize $P(\mathbb{R}^d \setminus \text{supp}(u_n))$ and small in order to minimize $U_n(\mathbb{R}^d \setminus \text{supp}(p))$ (since if g_n vanishes faster, $\lambda(\text{supp}(u_n) \setminus \text{supp}(p))$ decreases faster as well). This results in a simple heuristic, namely to set g_n to the smallest value such that $n_1 \leq t$, for some threshold t. For example, if we know the sample is "clean" in the sense that all points are drawn i.i.d. according to P, then we can take $t = 0$. A larger value of t could be chosen were we to suspect that the sample may contain noise, for

example $t = \log n$. In general, it remains an open question of how g_n should be optimized to balance between $P(\mathbb{R}^d \setminus \text{supp}(u_n))$ and $U_n(\mathbb{R}^d \setminus \text{supp}(p))$.

Remark 42. Cuevas and Fraiman (1997) use a plug-in approach to support estimation that can be leveraged here to further decrease $U_n(\mathbb{R}^d \setminus \text{supp}(p))$ when p has compact support and is continuously differentiable. Let $g_n = cn^{-\frac{1}{d+2}}$ for some constant c, and let α_n be such that $\alpha_n^{-1} = o\left(g_n^{-1}\right)$. For example, $\alpha_n \triangleq \sqrt{g_n}$, or if $d = o\left(\frac{\log n}{\log \log n}\right)$, then $\alpha_n \triangleq \frac{1}{\log n}$. Then, let G_P^n only contain the "lower-left" corners of grid cells containing more than $n\alpha_n$ sample points. Since this only decreases $U_n(\mathbb{R}^d \setminus \text{supp}(p))$, Lemma 32 remains correct and $U_n(\mathbb{R}^d \setminus \text{supp}(p)) \xrightarrow{\text{a.s.}} 0$. Furthermore, $\lambda(\text{supp}(p) \,\Delta\, \text{supp}(u_n)) \xrightarrow{\text{a.s.}} 0$ (Cuevas & Fraiman, 1997), and thus $P(\mathbb{R}^d \setminus \text{supp}(u_n)) \xrightarrow{\text{a.s.}} 0$, as well. One may use results given by Robert and Schapire (2000) to obtain an upper bound on $P(\mathbb{R}^d \setminus \text{supp}(u_n))$ for finite sample sizes.

Remark 43. It may be possible to improve on the convergence rate for the quantile-estimator, by using more information than 1 to 2 order statistics. This carries with it the risk of being less robust to classifier error. One such method is kernel-based quantile regression (Christmann & Steinwart, 2008), which is provably consistent. More complex quantile estimation methods may be useful in improving the convergence rate, without affecting the overall time complexity (dependent on the soft-classifier's time complexity), but these may exclude the use of ranking algorithms, as the quantile estimation method may rely on more than the relative ordering of the sample points.

7.4 Discussion

We have provided a computationally simple and consistent procedure for determining a δ-maximal level set estimator, which for measures that don't have a δ-jump, is also a δ-tight LDRF. While we have generated a uniform distribution for identifying low-density areas of P, this is not strictly necessary. Indeed, to return to the investment analogy, it is only necessary that the low-density areas have greater ROI than the high density areas. We term distributions which meet this condition as *lenient adversarial distributions*. The soft-classification approach used in this section applies for any such lenient adversarial distribution. Indeed, lenient adversarial distributions can also be used when the underlying mechanism is a hard-classifier. See (Nisenson, 2010) for a full discussion on lenient adversarial distributions and their relation to the existing SCC literature. The importance of these results, including the generation of a "tight" lenient adversarial distribution as given by the algorithm, lies not only in their justification of various approaches in the literature, but also in their applicability. Their only requirements are on the loss function used and that P be absolutely continuous with respect to the Lebesgue measure. Since most common loss functions satisfy the requirements and the condition on P is quite weak, a large body of results for regression and two-class classification can be utilized.

8. On The Dual SCC Problem

In the dual SCC problem the learner would like to guarantee the type II error, and minimize the type I error. This problem can be relevant to intrusion detection and authentication applications as well as to data mining and novelty detection. For example, in a biometric passport authentication system the authorities may mandate a maximal intruder pass rate. Under this constraint one would clearly want to minimize the false alarm rate. An alternative example is spam detection. A user may already have a two-class classification spam detection system in place. This system may perform very well at detecting spam which is similar to previously encountered spam. However, spammers are continually updating their spam so it will evade these filters. A second level system could be created, where an SCC classifier is trained on the legitimate e-mails. Any e-mails which the first-level determines as legitimate would then be tested against the second-level SCC classifier, which would either accept or reject them. A user may be willing to tolerate a certain level of spam from this second-level system, such as 1 in every 100 messages belonging to a new spam class getting through, but given that rate, would like as few legitimate messages as possible to be rejected.

Let $\delta_{\mathcal{Q}}$ be the maximally allowed type II error. Then the dual SCC problem is:

$$\begin{aligned} &\operatorname*{argmin}_{r} \rho(r, P) \\ &\text{such that: } \rho(r, Q) \geq 1 - \delta_{\mathcal{Q}}, \ \ \forall Q \in \mathcal{Q}, \end{aligned}$$

where $r(\cdot)$ is any function $\Omega \to [0, 1]$. When Ω is discrete and finite, this problem has a finite number of variables and a possibly infinite number of constraints depending on $\mathcal{Q}$. Thus, it is a linear semi-infinite program.

We represent by $r_I^*(\cdot)$ a solution to the primal problem, and by $r_{II}^*(\cdot)$ a solution to the dual problem. Define $\delta^* \stackrel{\triangle}{=} \rho(r_{II}^*, P)$ and $\delta_{\mathcal{Q}}^* \stackrel{\triangle}{=} 1 - \min_{Q \in \mathcal{Q}} \rho(r_I^*, Q)$. Since $r(\omega) \equiv \delta$ and $r(\omega) \equiv \delta_{\mathcal{Q}}$ are respectively feasible solutions to the primal and dual problems, $\delta^* \leq \delta_{\mathcal{Q}}$ and $\delta \leq \delta_{\mathcal{Q}}^*$.

Lemma 44. Let Ω be finite and discrete. If $\delta_{\mathcal{Q}}^* > 0$, then $\rho(r_I^*, P) = \delta$. If $\delta^* > 0$, then $\min_Q \rho(r_{II}^*, Q) = 1 - \delta_{\mathcal{Q}}$.

Proof Let $\delta_{\mathcal{Q}}^* > 0$. Let us assume by contradiction that $\rho(r_I^*, P) < \delta$. Then, define $r'(\omega) \stackrel{\triangle}{=} \min\left\{1, r_I^*(\omega) + \frac{\delta - \rho(r_I^*, P)}{N}\right\}$. Clearly, $\rho(r', P) \leq \delta$ and $\min_Q \rho(r', Q) > \min_Q \rho(r_I^*, Q)$. Contradiction.

Let $\delta^* > 0$. Let us assume by contradiction that $\min_Q \rho(r_{II}^*, Q) > 1 - \delta_{\mathcal{Q}}$. Then, define $r''(\omega) \stackrel{\triangle}{=} \max\left\{0, r_{II}^*(\omega) - \frac{\min_Q \rho(r_{II}^*, Q) - (1 - \delta_{\mathcal{Q}})}{N}\right\}$. Clearly, $\min_Q \rho(r'', Q) \geq 1 - \delta_{\mathcal{Q}}$ and $\rho(r'', P) < \rho(r_{II}^*, P)$. Contradiction. ■

We define $R_\delta^I \triangleq R_\delta^*$ as the set of primal-optimal rejection functions, and $R_{\delta_\mathcal{Q}}^{II}$ as the set of dual-optimal rejection functions. Examining the dual SCC problem in the investment analogy, the learner is assigned a target amount of money, $1-\delta_\mathcal{Q}$, which must be obtained on selling all assets. The learner's goal is to achieve this with the minimal starting investment. We can see that if the learner invests no money, then the amount of money made will fall short of the target. By investing in assets with higher ROI, the learner makes the most amount of progress towards the target with the least amount of money invested. Thus, we can see that the optimal investment strategy is likely to be similar to that of the primal problem. In fact, as shown by the following theorem, under mild conditions, the two sets of optimal strategies are identical.

Theorem 45 (Primal-Dual Equivalence). Let Ω be finite and discrete. If $\delta > 0$ and $\delta_\mathcal{Q}^* > 0$, then $R_{\delta_\mathcal{Q}^*}^{II} = R_\delta^I$. If $\delta_\mathcal{Q} > 0$ and $\delta^* > 0$, then $R_{\delta^*}^I = R_{\delta_\mathcal{Q}}^{II}$.

Proof Let $\delta > 0$ and $\delta_\mathcal{Q}^* > 0$. By Lemma 44, $\rho(r_I^*, P) = \delta$. Clearly, r_I^* is a feasible solution to the dual problem with $\delta_\mathcal{Q} = \delta_\mathcal{Q}^*$. Thus, $\delta^* \leq \delta$. Let us assume by contradiction that $\delta^* < \delta$. Then, there must exist some $r^*(\cdot)$ such that $\min_Q \rho(r^*, Q) \geq 1 - \delta_\mathcal{Q}^*$ and $\rho(r^*, P) < \delta$. Define $r'(\omega) \triangleq \min\left\{1, r^*(\omega) + \frac{\delta - \rho(r^*,P)}{N}\right\}$. Then, clearly $\rho(r', P) \leq \delta$, but $\min_Q \rho(r', Q) > \min_Q \rho(r^*, Q) \geq 1 - \delta_\mathcal{Q}^* = \min_Q \rho(r_I^*, Q)$. Contradiction. Therefore, $\delta^* = \delta$.

Since $\delta^* = \delta > 0$, by Lemma 44, $\min_Q \rho(r_{II}^*, Q) = 1 - \delta_\mathcal{Q}$. Thus, $r \in R_\delta^I$ if $\min_Q \rho(r, Q) = 1 - \delta_\mathcal{Q}^*$ and $\rho(r, P) = \delta$. Likewise, $r \in R_{\delta_\mathcal{Q}^*}^{II}$ if $\min_Q \rho(r, Q) = 1 - \delta_\mathcal{Q}^*$ and $\rho(r, P) = \delta$. Therefore, $R_{\delta_\mathcal{Q}^*}^{II} = R_\delta^I$.

Let $\delta_\mathcal{Q} > 0$ and $\delta^* > 0$. By Lemma 44, $\min_Q \rho(r_{II}^*, Q) = 1 - \delta_\mathcal{Q}$. Clearly, r_{II}^* is a feasible solution to the primal problem with $\delta = \delta^*$. Thus, $\delta_\mathcal{Q}^* \leq \delta_\mathcal{Q}$. Let us assume by contradiction that $\delta_\mathcal{Q}^* < \delta_\mathcal{Q}$. Then, there must exist some $r^*(\cdot)$ such that $\min_Q \rho(r^*, Q) > 1 - \delta_\mathcal{Q}$ and $\rho(r^*, P) \leq \delta^*$. Define $r''(\omega) \triangleq \max\left\{0, r^*(\omega) - \frac{\min_Q \rho(r^*,Q) - (1-\delta_\mathcal{Q})}{N}\right\}$. Then, clearly $\min_Q \rho(r'', Q) \geq 1 - \delta_\mathcal{Q}$, but $\rho(r'', P) < \rho(r^*, P) \leq \delta^* = \rho(r_{II}^*, P)$. Contradiction. Therefore, $\delta_\mathcal{Q}^* = \delta_\mathcal{Q}$.

Since $\delta_\mathcal{Q}^* = \delta_\mathcal{Q} > 0$, by Lemma 44, $\rho(r_I^*, P) = \delta = \delta^*$. Thus, $r \in R_\delta^I$ if $\min_Q \rho(r, Q) = 1 - \delta_\mathcal{Q}$ and $\rho(r, P) = \delta^*$. Likewise, $r \in R_{\delta_\mathcal{Q}}^{II}$ if $\min_Q \rho(r, Q) = 1 - \delta_\mathcal{Q}$ and $\rho(r, P) = \delta^*$. Therefore, $R_{\delta^*}^I = R_{\delta_\mathcal{Q}}^{II}$. ■

Using Theorem 45, it is trivial to solve the dual SCC problem where $\mathcal{Q} = \mathcal{Q}_\Lambda = \{Q : D_P(Q) \geq \Lambda\}$. To begin with, since we assume that $p_i > 0$ for all i, note that $\delta_\mathcal{Q} < 1 \Rightarrow \delta^* > 0$. Therefore, since $\delta_\mathcal{Q} > 0$ the optimal solution sets are identical by Theorem 45, and all the intermediate results, including Theorem 19, are correct when solving the primal problem

with $\delta = \delta^* > 0$. Therefore, it is trivial to construct a dual-analogue to Theorem 23. We also prove the analogue to Lemma 24.

Theorem 46 (Dual SCC Linear Program). An optimal soft rejection function and the optimal type I error, z_I, is obtained by solving the following linear program:

$$\begin{aligned} \text{minimize}_{r_1,r_2,\ldots,r_K,z_I} \;\; & z_I, \text{ subject to:} \\ & \sum_{i=1}^{K} r_i |S_i| p(S_i) \leq z_I \\ & 1 \geq r_1 \geq r_2 \geq \cdots \geq r_K \geq 0 \\ & r_w \geq 1 - \delta_{\mathcal{Q}} \\ & \rho_i \geq 1 - \delta_{\mathcal{Q}}, \; i \in \{1, 2, \ldots, |\mathcal{L}||\mathcal{H}| + |\mathcal{M}|\}. \end{aligned} \tag{3}$$

Lemma 47. Let r^* be the solution to the linear program. If r^* is vulnerable, then $r^* = r^{1-\delta_{\mathcal{Q}}}$.

Proof Let r^* be a vulnerable solution to the linear program (3), which clearly satisfies $1 \geq r_1^* \geq r_2^* \geq \cdots \geq r_K^* \geq 0$. Therefore, for all $i \in I_{min}(r^*)$, $r^*(i) = r_K^*$. We define z_I^* to be the minimal value of z_I that the linear program achieves for r^*. Let $j = \text{argmin}_{i \in I_{min}(r^*)} p_i$ and let S_u be the level set to which j belongs. We now prove that $u = 1$.

We first deal with the case where $D_P^{S_K} \geq \Lambda$ (in which case the constraint is completely vacuous). We note in this case that $S_1, S_2, \ldots, S_K \in \mathcal{L} \bigcup \mathcal{M}$, and therefore $w = K$. Thus, we have $r_1^* \geq r_2^* \geq \cdots \geq r_K^* = r_w^* \geq 1 - \delta_{\mathcal{Q}}$. Therefore, $\sum_{i=1}^{K} |S_i| p(S_i) r_i^* \geq r_K^* \geq 1 - \delta_{\mathcal{Q}}$. Therefore, $z_I^* \geq 1 - \delta_{\mathcal{Q}}$. We note that $r_1 = r_2 = \ldots r_K = 1 - \delta_{\mathcal{Q}}$ is a valid solution to the linear program for which $z_I = 1 - \delta_{\mathcal{Q}}$, which is the minimal value achievable. Therefore, $z_I^* = 1 - \delta_{\mathcal{Q}}$. If $u > 1$, then $r_1^* > r_K^* \geq z_I^* = 1 - \delta_{\mathcal{Q}}$ and $\sum_{i=1}^{K} |S_i| p(S_i) r_i^* > r_K^* \geq 1 - \delta_{\mathcal{Q}}$. Therefore, if $D_P^{S_K} \geq \Lambda$, $u = 1$.

We now turn our attention to the case where $D_P^{S_K} < \Lambda$. If we assume by contradiction that $u > 1$, then $r_{u-1}^* > r_u^* = r_{u+1}^* = \cdots = r_K^*$. $D_P\left(X^{(j)}\right) \geq \Lambda$ and by our assumption, $D_P^{S_K} < \Lambda$, which implies that $S_K \in \mathcal{H}$. If $D_P\left(X^{(j)}\right) > \Lambda$, then there exists some l for which $j \in S_u \equiv S_l \in \mathcal{L}$. The rejection rate for the level-set pair, (l, K), is $r_K^* \geq 1 - \delta_{\mathcal{Q}}$. Otherwise, $D_P\left(X^{(j)}\right) = \Lambda$, and $j \in S_u \equiv S_m \in \mathcal{M}$, and we have a rejection rate for S_m of $r_m^* = r_K^*$ and since $r_w \geq 1 - \delta_{\mathcal{Q}}$, this implies that $r_K^* = r_m^* = r_w \geq 1 - \delta_{\mathcal{Q}}$. Therefore, since $u > 1$, we get that $\sum_{i=1}^{K} |S_i| p(S_i) r_i^* > r_K^* \geq 1 - \delta_{\mathcal{Q}}$. Therefore, if $D_P^{S_K} < \Lambda$, $u = 1$.

Therefore, $u = 1$. This results in $r_1^* = r_2^* = \cdots = r_K^* \geq 1 - \delta_{\mathcal{Q}}$. Therefore, $\sum_{i=1}^{K} |S_i| p(S_i) r_i^* = r_K^*$ is clearly minimized by $r_K^* = 1 - \delta_{\mathcal{Q}}$, or $r^* = r^{1-\delta_{\mathcal{Q}}}$. ■

9. Concluding Remarks

We have introduced a game-theoretic approach to the SCC problem. In this approach the learner is opposed by an adversary. We believe that this viewpoint is essential for analyzing SCC applications such as intrusion detection and, in general, for "agnostic" analysis of single-class classification. This game-theoretic view lends itself well to analysis, allowing us to prove under what conditions low-density rejection is hard-optimal and if an optimal monotone rejection function is guaranteed to exist. Our analysis introduces soft decision strategies, which potentially allow for significantly better performance in our adversarial setting.

Observing the learner's futility when facing an omniscient and unlimited adversary, we considered restricted adversaries and provided full analysis of an interesting family of constrained games (in a decision-theoretic "Bayesian" setting where P is assumed to be known). The constraint we imposed on the adversary, given in terms of a divergence gap between the target and opposing distributions, is inspired by similar constraints used in "two-sample problem" related work in information theory (Ziv, 1988; Gutman, 1989; Ziv & Merhav, 1993). Of course, to compute the optimal learner strategy one has to know the exact value of this divergence gap, which is unknown in pure SCC problems. In applications we expect that something will be known or could be hypothesized about possible opposing distributions. For example, in biometric authentication, one should be able to statistically measure this gap. Thus, one could perhaps determine with a high confidence level that at least 99.9% of the population has a distribution with a KL-Divergence of at least 10 from the distribution of any member of the population. This can obviously be extended to k-factor authentication (see, e.g., Pointcheval & Zimmer, 2008). Assuming that the adversary may know $k-1$ factors, gaps can be found for each of the factors in order to ensure a particular intruder pass rate, with high probability. A different type of example occurs in extremely unbalanced two-class classification problems. Here, one could utilize the very few given examples from the other class to infer a bound on the gap. This complements the results of Kowalczyk and Raskutti (2003) where one-class learners were found to out-perform their two-class counterparts in some settings.

Our final major contribution is a simple and computationally feasible one-class classification algorithm. The SCC classifier is generated by thresholding a soft two-class classifier's output, where the output serves as a proxy for a density estimate, and a quantile estimate serves as the threshold. This approach can be extended to other use cases. For example, in (Yeh, Lee, & Lee, 2009), a multi-class classification problem is solved by constructing SVDD (D. Tax & Duin, 1999) one-class classifiers, where each class is described by a sphere, and learning a discriminant function which assigns a test point the class whose sphere center-point it is "nearest" to (the distance is normalized by various statistics). Instead of using

SVDD, our approach would be to create a two-class classifier for each class, where the second class is uniformly distributed over the active cells. A test point would be passed to each two-class classifier and the class chosen would be that belonging to the classifier which ranked the test point in the highest quantile (relative to the training sample for each class). This classification scheme makes sense because it labels the test point with the class for which it has the highest "relative" density (relative to other points within each class). Thus, we achieve the same goal without resorting to heuristics.

We have introduced a dual SCC problem and shown that, under very weak conditions, the solution sets for the primal and dual problems coincide. This allows one to easily extend results from one setting to the other, as we demonstrated by providing the dual solution to the constrained family of games considered earlier.

Various extension and generalizations to these results can be found in (Nisenson, 2010). These include extensions of Section 5 results to the infinite discrete setting and extensions to Section 7 giving additional results in the continuous setting such as a two-class reduction of SCC to hard binary-classification (as opposed to soft classification as we present here).

Our work can be extended in various ways and we believe that it opens up new avenues for future research and in particular could be useful for inspiring new algorithms for finite-sample SCC problems. One of the most important questions would be to determine convergence rates for the algorithm given in Section 7.3. It would be very nice to obtain an explicit expression for the lower bound output by the linear program of Theorem 23. Extensions of the analysis and algorithms for additional feature spaces, such as graphs or time-series, would be useful. An interesting question is whether performance, whether in terms of type II error or convergence rates, could be improved in different spaces. Clearly, the utilization of randomized strategies should be carried over to the finite sample case as well. A natural desirable extension is to extend our analysis for the soft setting to continuously infinite spaces.

We have focused in this work on "single-shot" games, meaning that the learner has to make a decision after every test observation. This is a very difficult setting as one cannot utilize cumulative statistics of the other class. Thus, we would expect that the results could be improved upon in a repeated-game setting, where several observations are provided from the same distribution, or in change point detection (Page, 1954; Hinkley, 1970), where one has to determine in a series of observations where the distribution P has been replaced by the (unknown) distribution Q as the underlying source. In the finite discrete setting, one should be able to easily extend some of the results here by replacing events with types (Cover & Thomas, 1991). Finally, a very interesting setting to consider is one in which the adversary has partial knowledge of the problem parameters and the learner's strategy. For example, the adversary may only know that P is in some subspace.

References

Bánhalmi, A., Kocsor, A., & Busa-Fekete, R. (2007). Counter-example generation-based one-class classification. In *Ecml '07: Proceedings of the 18th european conference on machine learning* (pp. 543–550). Berlin, Heidelberg: Springer-Verlag.

Bartlett, P. L., Jordan, M. I., & Mcauliffe, J. D. (2006, March). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*(473), 138–156.

Ben-David, S., & Lindenbaum, M. (1995). Learning distributions by their density-levels - a paradigm for learning without a teacher. In *Eurocolt* (pp. 53–68).

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image, and Signal Processing*, *141*(4), 217-222.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). Lof: Identifying density-based local outliers. In *Sigmod conference* (p. 93-104).

Cadre, B. (2006, April). Kernel estimation of density level sets. *Journal of Multivariate Analysis*, *97*(4), 999-1023. Available from `http://ideas.repec.org/a/eee/jmvana/v97y2006i4p999-1023.html`

Christmann, A., & Steinwart, I. (2008). Consistency of kernel-based quantile regression. *Appl. Stoch. Model. Bus. Ind.*, *24*(2), 171–183.

Clémençon, S., Lugosi, G., & Vayatis, N. (2005). Ranking and scoring using empirical risk minimization. In P. Auer & R. Meir (Eds.), *Colt* (Vol. 3559, p. 1-15). Springer.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. John Wiley & Sons.

Cuevas, A., & Fraiman, R. (1997). A plug-in approach to support estimation. *The Annals of Statistics*, *25*(6), 2300–2312.

Curry, R., & Heywood, M. I. (2009). One-class genetic programming. In *Eurogp '09: Proceedings of the 12th european conference on genetic programming* (pp. 1–12). Berlin, Heidelberg: Springer-Verlag.

Davenport, M. A., Baraniuk, R. G., & Scott, C. (2006). Learning minimum volume sets with support vector machines. In *Proc. ieee int. workshop on machine learning for signal processing (mlsp).*

Devroye, L., & Gyorfi, L. (2002). Distribution and density estimation. In L. Gyorfi (Ed.), *Principles of nonparametric learning* (pp. 190–201). Springer.

Devroye, L., & Wise, G. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, *38*, 480–488.

El-Yaniv, R., & Nisenson, M. (2006). Optimal single-class classification strategies. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Nips* (p. 377-384). MIT Press.

Grubbs, F. (1969, February). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1–21.

Gutman, M. (1989). Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. on Information Theory*, *35*(2), 401–408.

Hall, P., & Hannan, E. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika*, *75*(4), 705–714.

Hempstalk, K., Frank, E., & Witten, I. H. (2008). One-class classification by combining density and class probability estimation. In *Ecml pkdd '08: Proceedings of the 2008 european conference on machine learning and knowledge discovery in databases - part i* (pp. 505–519). Berlin, Heidelberg: Springer-Verlag.

Hinkley, D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, *57*, 1–17.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Juszczak, P., Tax, D. M. J., Pekalska, E., & Duin, R. P. W. (2009). Minimum spanning tree based one-class classifier. *Neurocomput.*, *72*(7-9), 1859–1869.

Kowalczyk, A., & Raskutti, B. (2003). Exploring fringe settings of svms for classification. In N. Lavrac, D. Gamberger, H. Blockeel, & L. Todorovski (Eds.), *Pkdd* (Vol. 2838, p. 278-290). Springer.

Lanckriet, G., Ghaoui, L. E., & Jordan, M. (2002). Robust novelty detection with single-class mpm. In *Nips* (p. 905-912).

Lanckriet, G. R. G., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*, 2002.

Lavrac, N., Gamberger, D., Blockeel, H., & Todorovski, L. (Eds.). (2003). *Knowledge discovery in databases: Pkdd 2003, 7th european conference on principles and practice of knowledge discovery in databases, cavtat-dubrovnik, croatia, september 22-26, 2003, proceedings* (Vol. 2838). Springer.

Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Sdm.*

Markou, M., & Singh, S. (2003a). Novelty detection: a review – part 1: statistical approaches. *Signal Processing*, *83*(12), 2481–2497.

Markou, M., & Singh, S. (2003b). Novelty detection: a review – part 2: neural network based approaches. *Signal Processing*, *83*(12), 2499–2521.

Minter, T. (1975). Single-class classification. In *Symposium on machine processing of remotely sensed data* (pp. 2A12–2A15).

Molchanov, I. S. (1990). Empirical estimation of distribution quantiles of random closed sets. *Theory of Probability and its Applications*, *35*(3), 594-600.

Nisenson, M. (2010). *On the foundations of adversarial single-class classification.* Unpublished master's thesis, Technion - Israel Institute of Technol-

ogy. Available from "http://www.cs.technion.ac.il/users/wwwb/cgi-bin/tr-get.cgi/2010/MSC/MSC-2010-18.pdf"

Nisenson, M., Yariv, I., El-Yaniv, R., & Meir, R. (2003). Towards behaviometric security systems: Learning to identify a typist. In N. Lavrac, D. Gamberger, H. Blockeel, & L. Todorovski (Eds.), *Pkdd* (Vol. 2838, p. 363-374). Springer.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1/2), 100–115.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076.

Pointcheval, D., & Zimmer, S. (2008). Multi-factor authenticated key exchange. In S. M. Bellovin, R. Gennaro, A. D. Keromytis, & M. Yung (Eds.), *Acns* (Vol. 5037, p. 277-295).

Rätsch, G., Mika, S., Schölkopf, B., & Müller, K. (2002). Constructing boosting algorithms from svms: An application to one-class classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, *24*(9), 1184–1199.

Rigollet, P., & Vert, R. (2008, October). *Optimal rates for plug-in estimators of density level sets.* Available from http://prunel.ccsd.cnrs.fr/docs/00/33/12/61/PDF/RigVert07_Bern3.pdf

Robert, D. M., & Schapire, R. E. (2000). On the convergence rate of good-turing estimators. In *In proceedings of the thirteenth annual conference on computational learning theory* (pp. 1–6). Morgan Kaufmann.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, *18*(3), 434–438.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443-1471.

Scott, C. (2007). Performance measures for neyman-pearson classification. *IEEE Transactions on Information Theory*, *53*(8), 2852-2863.

Scott, C. D., & Nowak, R. D. (2006). Learning minimum volume sets. *Journal of Machine Learning Research*, *7*, 665–704.

Steinwart, I., Hush, D., & Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, *6*, 211–232.

Steinwart, I., Hush, D. R., & Scovel, C. (2004). Density level detection is classification. In *Nips.*

Tax, D., & Duin, R. (1999). Support vector domain description. *Pattern Recognition Letter*, *20*(11-13), 1191–1199.

Tax, D. M. J., & Duin, R. P. W. (2001). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, *2*, 155-173.

Tsybakov, A. (1997). On nonparametric estimation of density level sets. *The Annals of Mathematical Statistics*, *25*(3), 948–969.

Uhlmann, W. (1963). Ranggrössen als schätzfunktionen. *Metrika*, *7*(1), 23–40.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vert, R., & Vert, J. (2006). Consistency and convergence rates of one-class svms and related algorithms. *J. Mach. Learn. Res.*, *7*, 817–854.

Yeh, C.-Y., Lee, Z.-Y., & Lee, S.-J. (2009). Boosting one-class support vector machines for multi-class classification. *Appl. Artif. Intell.*, *23*(4), 297–315.

Yu, H. (2005). Single-class classification with mapping convergence. *Machine Learning*, *61*(1-3), 49–69.

Zieliński, R. (2004, November). Optimal quantile estimators. small sample approach. *Technical report, Inst. of Math. Pol. Academy of Sci.*. Available from "http://www.impan.pl/Preprints/p653.pdf"

Ziv, J. (1988). On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, *34*, 278–286.

Ziv, J., & Merhav, N. (1993). A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, *39*(4), 1270–1279.

Appendix A. Section 5 Proofs

As a reminder, we assume that $p_i > 0$ for all $i \in \Omega$. Furthermore, for convenience we assume w.l.o.g. that Ω is defined such that $0 < p_1 \leq p_2 \leq \cdots \leq p_N$.

Lemma 48. Let $a + b = c + d$ and $a + c \geq b + d$. Then, $a \geq d$.

Proof Clearly $\frac{a}{2} \geq \frac{b+d-c}{2}$. Therefore, $a \geq \frac{a+b+d-c}{2} = d$ ■

Theorem 4 (Optimal Monotone Hard Decisions). When the learner is restricted to hard-decisions and $\mathcal{Q}$ satisfies Property A w.r.t. P, then there exists a monotone $r \in \mathcal{R}^*_\delta$.

Proof Recalling that $0 < p_1 \leq p_2 \leq \cdots \leq p_N$, we now define a rejection function as being x-monotone, if it is monotone up to index x. In other words, a rejection function, $r(\cdot)$ is x-monotone if $p_j < p_k \Rightarrow r(j) \geq r(k)$, for all $j < k \leq x$. Clearly, all rejection functions are 1-monotone, and a monotone rejection function is N-monotone.

Let us assume, by contradiction, that no monotone rejection function exists in $\mathcal{R}^*_\delta$. We will prove the existence of an N-monotone rejection function in $\mathcal{R}^*_\delta$ via induction. Let $r \in \mathcal{R}^*_\delta$. Then, r is $(k-1)$-monotone but not k-monotone, for some $2 \leq k \leq N$. Let $j = \min\{i : r(i) = 0\}$. We note that $1 \leq j < k$ and $r(k) = 1$ (otherwise, r would be k-monotone). We now prove the existence of a k-monotone rejection function, $r^* \in \mathcal{R}^*_\delta$. We define r^* as follows:

$$r^*(i) = \begin{cases} 1 & i = j, \\ 0 & i = k, \\ r(i) & \text{otherwise.} \end{cases}$$

Note that for all $i \leq j$, that $r^*(i) = 1$, and for all $j < i \leq k$, that $r^*(i) = 0$. Thus, r^* is a k-monotone rejection function. We now prove that $r^* \in \mathcal{R}^*_\delta$. Note that $\rho(r^*, P) = \rho(r, P) + p_j - p_k < \rho(r, P) \leq \delta$, and thus r^* is a δ-valid hard rejection function. Let $Q^* \in \mathcal{Q}$ be such that $\min_Q \rho(r^*, Q) = \rho(r^*, Q^*) = \rho(r, Q^*) + q^*_j - q^*_k$. Thus, if $q^*_j \geq q^*_k$, $\rho(r^*, Q^*) \geq \rho(r, Q^*)$. Otherwise, there exists $Q^{*\prime}$ as in Property A and in particular, by Lemma 48, $q^*_j \geq q^{*\prime}_k$. Consequently, $\rho(r^*, Q^*) = \rho(r, Q^{*\prime}) + q^*_j - q^{*\prime}_k \geq \rho(r, Q^{*\prime})$. Therefore, there always exists $Q \in \mathcal{Q}$ such that $\rho(r^*, Q^*) \geq \rho(r, Q)$ (either $Q = Q^*$ or $Q = Q^{*\prime}$). Therefore, $\min_Q \rho(r^*, Q) \geq \min_Q \rho(r, Q)$, and thus, $r^* \in \mathcal{R}^*_\delta$. Therefore, by induction, there must exist an optimal N-monotone rejection function. Contradiction. ■

Remark 49. The above proof works for a weaker version of Property A: If for all $p_j < p_k$ and $Q \in \mathcal{Q}$ for which $q_j < q_k$, there exists a distribution $Q' \in \mathcal{Q}$ such that $q'_j - q'_k + \sum_{i=1}^{j}(q_i -$

$q'_i) + \sum_{i=k+1} \min\{q_i - q'_i, 0\} \geq 0$. As used in the proof, this would read:

$$\begin{aligned}
\rho(r^*, Q^*) =& \rho(r^*, Q^{*\prime}) + \sum_{i=1}^{N} r^*(i)(q^*_i - {q^*}'_i) \\
=& \rho(r, Q^{*\prime}) + {q^*}'_j - {q^*}'_k + \sum_{i=1}^{j}(q^*_i - {q^*}'_i) + \sum_{i=k+1} r^*(i)(q^*_i - {q^*}'_i) \\
\geq& \rho(r, Q^{*\prime}) + {q^*}'_j - {q^*}'_k + \sum_{i=1}^{j}(q^*_i - {q^*}'_i) + \sum_{i=k+1} \min\{q^*_i - {q^*}'_i, 0\} \\
\geq& \rho(r, Q^{*\prime}).
\end{aligned}$$

Remark 50. If we strengthen the condition in Property A from $q_j + q'_j \geq q_k + q'_k$ to $q_j + q'_j > q_k + q'_k$ for all distributions Q such that $q_j \leq q_k$ (instead of $q_j < q_k$), then all optimal rejection functions would be monotone. Note that the set of all distributions does not have this modified property, but the set of all distributions bounded away from zero ($\{Q : q_i > 0, \forall i \in \Omega\}$) does.

Theorem 6 (Optimal Monotone Soft Decisions).
If $\mathcal{Q}$ satisfies Property B w.r.t. P, then there exists an optimal strictly monotone rejection function.

Proof We note that the condition for strict-monotonicity is equivalent to $p_j \leq p_k \Rightarrow r(j) \geq r(k)$, and that $0 < p_1 \leq p_2 \leq \cdots \leq p_N$. We now define an x-right-strictly-monotone rejection function as one which has strictly-monotone properties for the last x indices. In other words, a rejection function $r(\cdot)$ is x-right-strictly-monotone if $p_j \leq p_k \Rightarrow r(j) \geq r(k)$, for all $j < k$, $k > N - x$. Clearly, all rejection functions are 0-right-strictly-monotone, and an N-right-strictly monotone rejection function is strictly monotone.

We assume contradictorily that there is no such rejection function. Let $r \in \mathcal{R}^*_\delta$. We note that r is $(v-1)$-right-strictly-monotone but not v-right-strictly-monotone for some $1 \leq v \leq N$. We will prove by induction that there exists an N-right-strictly-monotone function in $\mathcal{R}^*_\delta$. Let $k = N - v + 1$. Since r is not v-right-strictly-monotone, then there must exist some $j < k$ for which $p_j \leq p_k$ and $r(j) < r(k)$. Define, for any event ω and distribution D:

$$\begin{aligned}
S_r(\omega) &\triangleq \{i : p_i = p_\omega \wedge r(i) = r(\omega)\}; \\
g(D, \omega) &\triangleq \frac{\sum_{i \in S_r(\omega)} d_i}{|S_r(\omega)| p_\omega} = \frac{\sum_{i \in S_r(\omega)} \frac{d_i}{p_\omega}}{|S_r(\omega)|}.
\end{aligned}$$

$S_r(\omega)$ is the intersection of ω's probability level-set with ω's rejection level-set. $g(D, \omega)$ is simply an average of the elements of D corresponding to symbols in $S_r(\omega)$ normalized by

$\frac{1}{p_\omega}$. We note that $g(P,\omega)=1$ always. We define r^* as follows:

$$r^*(i)=\begin{cases}\frac{|S_r(j)|p_j r(j)+|S_r(k)|p_k r(k)}{|S_r(j)|p_j+|S_r(k)|p_k} & i\in S_r(j)\cup S_r(k),\\ r(i) & \text{otherwise;}\end{cases}$$

$$\begin{aligned}
\Rightarrow r^*(j)-r(j) =& \frac{|S_r(j)|p_j r(j)+|S_r(k)|p_k r(k)}{|S_r(j)|p_j+|S_r(k)|p_k}-r(j)\\
=& \frac{|S_r(k)|p_k(r(k)-r(j))}{|S_r(j)|p_j+|S_r(k)|p_k}>0\\
r(k)-r^*(k) =& \frac{|S_r(j)|p_j(r(k)-r(j))}{|S_r(j)|p_j+|S_r(k)|p_k}=\frac{|S_r(j)|p_j}{|S_r(k)|p_k}(r^*(j)-r(j))\\
\Rightarrow \forall D,\rho(r^*,D)-\rho(r,D) =& \left[(r^*(j)-r(j))\sum_{i\in S_r(j)}d_i\right]+\left[(r^*(k)-r(k))\sum_{i\in S_r(k)}d_i\right]\\
=&(r^*(j)-r(j))\left[\sum_{i\in S_r(j)}d_i-\frac{|S_r(j)|p_j}{|S_r(k)|p_k}\sum_{i\in S_r(k)}d_i\right]\\
=&(r^*(j)-r(j))|S_r(j)|p_j\left[\frac{\sum_{i\in S_r(j)}d_i}{|S_r(j)|p_j}-\frac{\sum_{i\in S_r(k)}d_i}{|S_r(k)|p_k}\right]\\
=&(r^*(j)-r(j))|S_r(j)|p_j\left[g(D,j)-g(D,k)\right].
\end{aligned}$$

Therefore, noting that $r^*(j)>r(j)$,

$$\rho(r^*,D)<\rho(r,D)\Rightarrow\ g(D,j)<g(D,k)\Rightarrow\ \min_{i\in S_r(j)}\frac{d_i}{p_j}<\max_{i\in S_r(k)}\frac{d_i}{p_k}. \tag{4}$$

Since $g(P,j)=g(P,k)=1$, $\rho(r^*,P)=\rho(r,P)=\delta$. Therefore, r^* is a valid rejection function. Let $u>k$. We note by the definition of r^* and the fact that r is $(v-1)$-right-strictly-monotone that $r^*(u)=r(u)\le r(j)<r^*(j)=r^*(k)<r(k)$. Therefore, r^* is still $(v-1)$-right-strictly-monotone (but not necessarily v-right-strictly-monotone).

Let Q^* be such that $\rho(r^*,Q^*)=\min_Q\rho(r^*,Q)$. We will now show that $\exists\hat{Q}\in\mathcal{Q}$ s.t. $\rho(r^*,Q^*)\ge\rho(r,\hat{Q})$ (and therefore, $\min_Q\rho(r^*,Q)\ge\min_Q\rho(r,Q)$). The following algorithm finds such a $\hat{Q}$:

1. Set $Q=Q^*$.
2. while $\rho(r^*,Q)<\rho(r,Q)$
   (a) Let a and b be such that $q_a=\min_{i\in S_r(j)}q_i$ and $q_b=\max_{i\in S_r(k)}q_i$. We note that $\rho(r^*,Q)<\rho(r,Q)\Rightarrow\frac{q_a}{p_j}<\frac{q_b}{p_k}\Rightarrow\frac{q_a}{p_a}<\frac{q_b}{p_b}$.
   (b) Since $\mathcal{Q}$ satisfies Property B, there exists a $Q'\in\mathcal{Q}$ which is identical to Q for all $i\neq a,b$ and such that $\frac{q'_a}{p_a}\ge\frac{q'_b}{p_b}$. Set $Q=Q'$.

3. end while. Output $\hat{Q} = Q$.

Since for all iterations, $r^*(a) = r^*(b)$, at step (b) we have $\rho(r^*, Q') = \rho(r^*, Q) = \rho(r^*, Q^*)$. After setting $Q = Q'$ at step (b), we have $\frac{q_a}{p_a} \geq \frac{q_b}{p_b}$, and therefore the loop never repeats for the same pair of symbols (a, b). Therefore, the loop is guaranteed to terminate. After ending, $\rho(r^*, Q^*) = \rho(r^*, \hat{Q}) \geq \rho(r, \hat{Q})$, so $\min_Q \rho(r^*, Q) \geq \min_Q \rho(r, Q)$, and $r^* \in \mathcal{R}^*_\delta$.

While there still exists a j such that $r^*(j) < r^*(k)$ we relabel r^* as r and repeat the above procedure (note that it never repeats for the same pair j, k). The resulting r^* is $(v-1)$-right-strictly-monotone as shown above, but since now $j < k \Rightarrow r^*(j) \geq r^*(k)$, r^* is v-right-strictly-monotone.

Thus, by induction there exists an optimal N-right-strictly-monotone rejection function, which is a contradiction. ■

Remark 51. Strengthening the conditions in Property B to $\frac{q_j}{p_j} \leq \frac{q_k}{p_k}$ and $\frac{q'_j}{p_j} > \frac{q'_k}{p_k}$ would strengthen Theorem 6 so that all optimal rejection functions are strictly monotone. Once more, the set of all distributions does not have this modified property, but the set of all distributions bounded away from zero does.

Theorem 10 (LDRS optimality). Let r^* be an LDRF. Let r be any monotone δ-valid rejection function. Then

$$\min_{Q \in \mathcal{Q}} \rho(r^*, Q) \geq \min_{Q \in \mathcal{Q}} \rho(r, Q),$$

for any $\mathcal{Q}$ satisfying Property C. Thus, if $\mathcal{Q}$ possess both Property A and Property C w.r.t. P, then LDRS is hard-optimal.

Proof We define, for a hard rejection function r, $\theta(r) \stackrel{\triangle}{=} \min_{\omega: r(\omega)=0} p_\omega$, $Z_\theta(r) \stackrel{\triangle}{=} \{\omega : p_\omega = \theta(r) \wedge r(\omega) = 1\}$ and $z_\theta(r) \stackrel{\triangle}{=} |Z_\theta(r)|$.

Assume, by contradiction, that $\min_{Q \in \mathcal{Q}} \rho(r^*, Q) < \min_{Q \in \mathcal{Q}} \rho(r, Q)$. Let Q^* be the minimizer of $\rho(r^*, Q)$. Then, $\rho(r^*, Q^*) < \rho(r, Q^*)$. If $\theta(r) > \theta(r^*)$ then, by the definition of LDRF and by the monotonicity of r, $\rho(r, P) > \delta$, which contradicts r's validity. If $\theta(r) < \theta(r^*)$ then, by r's monotonicity, $r(\omega) = 1 \Rightarrow r^*(\omega) = 1$, and for any distribution D, $\rho(r, D) \leq \rho(r^*, D)$, contradicting $\rho(r^*, Q^*) < \rho(r, Q^*)$. Therefore, $\theta(r) = \theta(r^*)$. If $z_\theta(r) > z_\theta(r^*)$ then $\rho(r, P) > \delta$ since r^* is an LDRF. Otherwise, $z_\theta(r) \leq z_\theta(r^*)$, and by Property C the set $\mathcal{Q}$ contains all distributions identical to Q^* up to a permutation of the θ-probability events. Therefore, $\min_{Q \in \mathcal{Q}} \rho(r^*, Q) \geq \min_{Q \in \mathcal{Q}} \rho(r, Q)$. Contradiction. ■

Appendix B. Section 6 Proofs

Lemma 24. Let r^* be the solution to the linear program. If r^* is vulnerable, then $r^* = r^\delta$.

Proof Let r^* be a vulnerable solution to the linear program (2), which clearly satisfies $1 \geq r_1^* \geq r_2^* \geq \cdots \geq r_K^* \geq 0$. Therefore, for all $i \in I_{min}(r^*)$, $r^*(i) = r_K^*$. We define z^* to be the maximal value of z that the linear program achieves for r^*. Let $j = \text{argmin}_{i \in I_{min}(r^*)} p_i$ and let S_u be the level set to which j belongs. We now prove that $u = 1$.

We first deal with the case where $D_P^{S_K} \geq \Lambda$ (in which case the constraint is completely vacuous). We note in this case that $S_1, S_2, \ldots, S_K \in \mathcal{L} \bigcup \mathcal{M}$, and therefore $w = K$. Thus, we have $r_1^* \geq r_2^* \geq \cdots \geq r_K^* = r_w^* \geq z^*$. We note that $z^* \leq \delta$, otherwise $\delta = \sum_{i=1}^K |S_i| p(S_i) r_i^* \geq r_K^* \geq z^* > \delta$. We note that $r_1 = r_2 = \ldots r_K = \delta$ is a valid solution to the linear program for which $z = \delta$, which is the maximal value achievable. Therefore, $z^* = \delta$. If $u > 1$, then $r_1^* > r_K^* \geq z^* = \delta$ and $\sum_{i=1}^K |S_i| p(S_i) r_i^* > r_K^* \geq \delta$. Therefore, if $D_P^{S_K} \geq \Lambda$, $u = 1$.

We now turn our attention to the case where $D_P^{S_K} < \Lambda$. If we assume by contradiction that $u > 1$, then $r_{u-1}^* > r_u^* = r_{u+1}^* = \cdots = r_K^*$. We define $\Pr[S] \triangleq |S| p(S)$ for a level set S, and $c \triangleq \frac{\Pr[S_{u-1}]}{\sum_{i=u}^K \Pr[S_i]}$. Let $0 < \epsilon < \frac{r_{u-1}^* - r_u^*}{c+1}$. We now define a new rejection function r' as follows:

$$r_i' \triangleq \begin{cases} r_i^* & i < u-1, \\ r_i^* - \epsilon & i = u-1, \\ r_i^* + c\epsilon & i \geq u. \end{cases}$$

We note that:

$$\begin{aligned} \rho(r', P) =& \rho(r^*, P) - \Pr[S_{u-1}]\epsilon + \sum_{i=u}^K \Pr[S_i] c\epsilon \\ =& \rho(r^*, P) - \Pr[S_{u-1}]\epsilon + \Pr[S_{u-1}]\epsilon = \rho(r^*, P) = \delta. \end{aligned}$$

Therefore, r' is δ-valid. Let z' be the maximal value of z that the linear program achieves for r'. $D_P\left(X^{(j)}\right) \geq \Lambda$ and by our assumption, $D_P^{S_K} < \Lambda$, which implies that $S_K \in \mathcal{H}$. If $D_P\left(X^{(j)}\right) > \Lambda$, then there exists some l for which $j \in S_u \equiv S_l \in \mathcal{L}$. The rejection rate for the level-set pair, (l, K), is r_K^*. Otherwise, $D_P\left(X^{(j)}\right) = \Lambda$, and $j \in S_u \equiv S_m \in \mathcal{M}$, and we have a rejection rate for S_m of $r_m^* = r_K^*$. Since z^* cannot be less than $r_{min}^* = r_K^*$, we have $z^* = r_K^*$ in both cases ($r_K^* \geq z^* \geq r_K^*$). We note that:

$$r_{u-1}' - r_u' = (r_{u-1}^* - \epsilon) - (r_u^* + c\epsilon) = (r_{u-1}^* - r_u^*) - (c+1)\epsilon > 0$$

Clearly for $i > u$, $r_{i-1}' = r_i' = r_K^* + c\epsilon$. Obviously, $1 \geq r_1'$ and $r_K' > r_K^* \geq 0$. Therefore, $1 \geq r_1' \geq r_2' \geq \cdots \geq r_K' > 0$, and r' is a feasible solution to the linear program. Furthermore, $z' \geq r_{min}' = r_K' > r_K^* = z^*$, which contradicts the fact that r^* maximizes z (and is the solution to the linear program).

Therefore, $u = 1$. This results in $r_1^* = r_2^* = \cdots = r_K^*$. Since $\rho(r^*, P) = \delta$, we have that $\delta = \sum_{i=1}^K |S_i| p(S_i) r_i^* = r_K^*$, or $r^* = r^\delta$. ■

Appendix C. Section 7 Proofs

We begin by providing some additional definitions. Let $\mathbb{B}$ be the set of all Borel sets over $\mathbb{R}^d$. For two Borel sets a, b we define $a \overset{\lambda}{=} b \Leftrightarrow \lambda(a \,\Delta\, b) = 0$, where Δ is the symmetric difference operator. For two functions, f, g over $\mathbb{R}^d$ and Borel set b, define $\Delta_b(f,g) \overset{\triangle}{=} \{x \in b : f(x) \neq g(x)\}$. We define the function $\mathbb{I}_b(x) \overset{\triangle}{=} \mathbb{I}(x \in b)$, where $\mathbb{I}(\cdot)$ is the indicator function.

Lemma 52. Let $m' \in \text{core}_\delta(P)$. Let m be a Borel set such that $m \overset{\lambda}{=} m'$. Then $m \in \text{core}_\delta(P)$.

Proof Since $m' \in \text{core}_\delta(P)$, $P(m') = \delta$ and there exists a minimum volume set b' of measure $1-\delta$, such that $m' \bigcap b' = \emptyset$. Let $b \overset{\triangle}{=} b' \setminus m$. We note that $\lambda(m \,\Delta\, m') = 0$. Therefore, $b = b' \setminus m \overset{\lambda}{=} b' \setminus m' = b'$. Therefore, b' is a minimum volume set of measure $1-\delta$. Since $m \bigcap b = \emptyset$ and $P(m) = \delta$, $m \in \text{core}_\delta(P)$. ■

Theorem 30 (LDRS optimality - Continuous Setting). When the learner is restricted to hard-decisions and $\mathcal{Q}$ satisfies Property A_{cont} w.r.t. P, then LDRS is optimal.

Proof Assume that the statement is false. Therefore, there must exist some $m \in \text{core}_\delta(P)$ such that for all $r \in R^*_\delta$, $\lambda(\Delta_{\mathbb{R}^d}(r, \mathbb{I}_{l_P(m)})) > 0$. Let $r' \in R^*_\delta$, such that $\rho(r', P) = \delta$. Define

$$r(x) \overset{\triangle}{=} \begin{cases} 1 & p(x) = 0, \\ r'(x) & \text{otherwise.} \end{cases}$$

Therefore, $r \in R^*_\delta$ and $\lambda(\Delta_{\mathbb{R}^d}(r, \mathbb{I}_{l_P(m)})) > 0$. Let $j = \{x \in m : r(x) = 0\}$. Therefore, $P(j) > 0$. Thus, there must exist a set k, such that $k \bigcap m = \emptyset$, $k \subset \text{supp}(p)$, $\int_k r(x)\lambda(dx) = \lambda(k)$, and $P(k) = P(j)$ (otherwise, $\rho(r', P) \neq \delta$). Since $P(k) = P(j) > 0$ and $m \in \text{core}_\delta(P)$, we have $\lambda(j) \geq \lambda(k)$. We define:

$$r^*(x) \overset{\triangle}{=} \begin{cases} 1 & x \in j, \\ 0 & x \in k, \\ r(x) & \text{otherwise.} \end{cases}$$

We note that $\rho(r^*, P) = \rho(r, P) \leq \delta$.

Let $Q^* \in \mathcal{Q}$ be such that $\min_Q \rho(r^*, Q) = \rho(r^*, Q^*) = \rho(r, Q^*) + Q^*(j) - Q^*(k)$. Thus, if $Q^*(j) \geq Q^*(k)$, $\rho(r^*, Q^*) \geq \rho(r, Q^*)$. Otherwise, there exists $Q^{*\prime}$ as in Property A_{cont} and in particular, by Lemma 48, $Q^*(j) > Q^{*\prime}(k)$. Consequently, $\rho(r^*, Q^*) = \rho(r, Q^{*\prime}) + Q^*(j) - Q^*(k) \geq \rho(r, Q^{*\prime})$. Therefore, there always exists $Q \in \mathcal{Q}$ such that $\rho(r^*, Q^*) \geq \rho(r, Q)$ (either $Q = Q^*$ or $Q = Q^{*\prime}$). Therefore, $\min_Q \rho(r^*, Q) \geq \min_Q \rho(r, Q)$, and thus $r^* \in R^*_\delta$. However, $\lambda(\Delta_{\mathbb{R}^d}(r^*, \mathbb{I}_{l_P(m)})) = 0$. Contradiction. ■